

Marina Gavrilova et al. (Eds.)

LNCS 3982

Computational Science and Its Applications – ICCSA 2006

International Conference
Glasgow, UK, May 2006
Proceedings, Part III

3
Part III

 Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

University of Dortmund, Germany

Madhu Sudan

Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Moshe Y. Vardi

Rice University, Houston, TX, USA

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Marina Gavrilova Osvaldo Gervasi
Vipin Kumar C.J. Kenneth Tan
David Taniar Antonio Laganà
Youngsong Mun Hyunseung Choo (Eds.)

Computational Science and Its Applications – ICCSA 2006

International Conference
Glasgow, UK, May 8-11, 2006
Proceedings, Part III

Volume Editors

Marina Gavrilova
University of Calgary, Canada
E-mail: marina@cpsc.ucalgary.ca

Osvaldo Gervasi
University of Perugia, Italy
E-mail: ogervasi@computer.org

Vipin Kumar
University of Minnesota, Minneapolis, USA
E-mail: kumar@cs.umn.edu

C.J. Kenneth Tan
OptimaNumerics Ltd., Belfast, UK
E-mail: cjtan@optimanumerics.com

David Taniar
Monash University, Clayton, Australia
E-mail: david.taniar@infotech.monash.edu.au

Antonio Laganà
University of Perugia, Italy
E-mail: lag@unipg.it

Youngsong Mun
SoongSil University, Seoul, Korea
E-mail: mun@computing.soongsil.ac.kr

Hyunseung Choo
Sungkyunkwan University, Suwon, Korea
E-mail: choo@ece.skku.ac.kr

Library of Congress Control Number: 2006925086

CR Subject Classification (1998): F, D, G, H, I, J, C.2-3

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN 0302-9743
ISBN-10 3-540-34075-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-34075-1 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springer.com

© Springer-Verlag Berlin Heidelberg 2006
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11751595 06/3142 5 4 3 2 1 0

Preface

This five-volume set was compiled following the 2006 International Conference on Computational Science and its Applications, ICCSA 2006, held in Glasgow, UK, during May 8–11, 2006. It represents the outstanding collection of almost 664 refereed papers selected from over 2,450 submissions to ICCSA 2006.

Computational science has firmly established itself as a vital part of many scientific investigations, affecting researchers and practitioners in areas ranging from applications such as aerospace and automotive, to emerging technologies such as bioinformatics and nanotechnologies, to core disciplines such as mathematics, physics, and chemistry. Due to the sheer size of many challenges in computational science, the use of supercomputing, parallel processing, and sophisticated algorithms is inevitable and becomes a part of fundamental theoretical research as well as endeavors in emerging fields. Together, these far-reaching scientific areas contributed to shaping this conference in the realms of state-of-the-art computational science research and applications, encompassing the facilitating theoretical foundations and the innovative applications of such results in other areas.

The topics of the refereed papers span all the traditional as well as emerging computational science realms, and are structured according to the five major conference themes:

- Computational Methods, Algorithms and Applications
- High-Performance Technical Computing and Networks
- Advanced and Emerging Applications
- Geometric Modeling, Graphics and Visualization
- Information Systems and Information Technologies

Moreover, submissions from 31 workshops and technical sessions in areas such as information security, mobile communication, grid computing, modeling, optimization, computational geometry, virtual reality, symbolic computations, molecular structures, Web systems and intelligence, spatial analysis, bioinformatics and geocomputations, are included in this publication. The continuous support of computational science researchers has helped ICCSA to become a firmly established forum in the area of scientific computing.

We recognize the contribution of the International Steering Committee and sincerely thank the International Program Committee for their tremendous support in putting this conference together, the near 800 referees for their diligent work, and the IEE European Chapter for their generous assistance in hosting the event.

We also thank our sponsors for their continuous support without which this conference would not be possible.

Finally, we thank all authors for their submissions and all invited speakers and conference attendants for making the ICCSA Conference truly one of the premium events on the scientific community scene, facilitating exchange of ideas, fostering new collaborations, and shaping the future of computational science.

May 2006

Marina L. Gavrilova
Oswaldo Gervasi

on behalf of the co-editors
Vipin Kumar
Chih Jeng Kenneth Tan
David Taniar
Antonio Laganà
Youngsong Mun
Hyunseung Choo

Organization

ICCSA 2006 was organized by the Institute of Electrical Engineers (IEE)(UK), the University of Perugia (Italy), Calgary University (Canada) and Minnesota University (USA).

Conference Chairs

Vipin Kumar (University of Minnesota, Minneapolis, USA), Honorary Chair
Marina L. Gavrilova (University of Calgary, Calgary, Canada), Conference
Co-chair, Scientific
Osvaldo Gervasi (University of Perugia, Perugia, Italy), Conference Co-chair,
Program

Steering Committee

Vipin Kumar (University of Minnesota, USA)
Marina L. Gavrilova (University of Calgary, Canada)
Osvaldo Gervasi (University of Perugia, Perugia, Italy)
C. J. Kenneth Tan (OptimaNumerics, UK)
Alexander V. Bogdanov (Institute for High Performance Computing
and Data Bases, Russia)
Hyunseung Choo (Sungkyunkwan University, Korea)
Andres Iglesias (University of Cantabria, Spain)
Antonio Laganà (University of Perugia, Italy)
Heow-Pueh Lee (Institute of High Performance Computing, Singapore)
Youngsong Mun (Soongsil University, Korea)
David Taniar (Monash University, Australia)

Workshop Organizers

Applied Cryptography and Information Security (ACIS 2006)

Sherman S.M. Chow (New York University, USA)
Joseph K. Liu (University of Bristol, UK)
Patrick Tsang (Dartmouth College, USA)
Duncan S Wong (City University of Hong Kong, Hong Kong)

Approaches or Methods of Security Engineering (AMSE 2006)

Haeng Kon Kim (Catholic University of Daegu, Korea)
Tai-hoon Kim (Korea Information Security Agency, Korea)

Authentication, Authorization and Accounting (AAA 2006)
Haeng Kon Kim (Catholic University of Daegu, Korea)

Computational Geometry and Applications (CGA 2006)
Marina Gavrilova (University of Calgary, Calgary, Canada)

Data Storage Devices and Systems (DSDS 2006)
Yeonseung Ryu (Myongji University, Korea)
Junho Shim (Sookmyong Womens University, Korea)
Youjip Won (Hanyang University, Korea)
Yongik Eom (Seongkyunkwan University, Korea)

Embedded System for Ubiquitous Computing (ESUC 2006)
Tei-Wei Kuo (National Taiwan University, Taiwan)
Jiman Hong (Kwangwoon University, Korea)

4th Technical Session on Computer Graphics (TSCG 2006)
Andres Iglesias (University of Cantabria, Spain)
Deok-Soo Kim (Hanyang University, Korea)

GeoComputation (GC 2006)
Yong Xue (London Metropolitan University, UK)

Image Processing and Computer Vision (IPCV 2006)
Jiawan Zhang (Tianjin University, China)

**Intelligent Services and the Synchronization in Mobile
Multimedia Networks (ISS 2006)**
Dong Chun Lee (Howon University, Korea)
Kuinam J Kim (Kyonggi University, Korea)

**Integrated Analysis and Intelligent Design Technology
(IAIDT 2006)**
Jae-Woo Lee (Konkuk University, Korea)

Information Systems Information Technologies (ISIT 2006)
Youngsong Mun (Soongsil University, Korea)

Information Engineering and Applications in Ubiquitous Computing Environments (IEAUCE 2006)

Sangkyun Kim (Yonsei University, Korea)

Hong Joo Lee (Dankook University, Korea)

Internet Communications Security (WICS 2006)

Sierra-Camara José Maria (University Carlos III of Madrid, Spain)

Mobile Communications (MC 2006)

Hyunseung Choo (Sungkyunkwan University, Korea)

Modelling Complex Systems (MCS 2006)

John Burns (Dublin University, Ireland)

Ruili Wang (Massey University, New Zealand)

Modelling of Location Management in Mobile Information Systems (MLM 2006)

Dong Chun Lee (Howon University, Korea)

Numerical Integration and Applications (NIA 2006)

Elise de Doncker (Western Michigan University, USA)

Specific Aspects of Computational Physics and Wavelet Analysis for Modelling Suddenly-Emerging Phenomena in Nonlinear Physics, and Nonlinear Applied Mathematics (PULSES 2006)

Carlo Cattani (University of Salerno, Italy)

Cristian Toma (Titu Maiorescu University, Romania)

Structures and Molecular Processes (SMP 2006)

Antonio Laganà (University of Perugia, Perugia, Italy)

Optimization: Theories and Applications (OTA 2006)

Dong-Ho Lee (Hanyang University, Korea)

Deok-Soo Kim (Hanyang University, Korea)

Ertugrul Karsak (Galatasaray University, Turkey)

Parallel and Distributed Computing (PDC 2006)

Jiawan Zhang (Tianjin University, China)

Pattern Recognition and Ubiquitous Computing (PRUC 2006)

Jinok Kim (Daegu Haany University, Korea)

Security Issues on Grid/Distributed Computing Systems (SIGDCS 2006)

Tai-Hoon Kim (Korea Information Security Agency, Korea)

Technologies and Techniques for Distributed Data Mining (TTDDM 2006)

Mark Baker (Portsmouth University, UK)

Bob Nichol (Portsmouth University, UK)

Ubiquitous Web Systems and Intelligence (UWSI 2006)

David Taniar (Monash University, Australia)

Eric Pardede (La Trobe University, Australia)

Ubiquitous Application and Security Service (UASS 2006)

Yeong-Deok Kim (Woosong University, Korea)

Visual Computing and Multimedia (VCM 2006)

Abel J. P. Gomes (University Beira Interior, Portugal)

Virtual Reality in Scientific Applications and Learning (VRSAL 2006)

Oswaldo Gervasi (University of Perugia, Italy)

Antonio Riganelli (University of Perugia, Italy)

Web-Based Learning (WBL 2006)

Woochun Jun Seoul (National University of Education, Korea)

Program Committee

Jemal Abawajy (Deakin University, Australia)
Kenny Adamson (EZ-DSP, UK)
Srinivas Aluru (Iowa State University, USA)
Mir Atiqullah (Saint Louis University, USA)
Frank Baetke (Hewlett Packard, USA)
Mark Baker (Portsmouth University, UK)
Young-Cheol Bang (Korea Polytechnic University, Korea)
David Bell (Queen's University of Belfast, UK)
Stefania Bertazzon (University of Calgary, Canada)
Sergei Bepamyatnikh (Duke University, USA)
J. A. Rod Blais (University of Calgary, Canada)
Alexander V. Bogdanov (Institute for High Performance Computing
and Data Bases, Russia)
Peter Brezany (University of Vienna, Austria)
Herve Bronnimann (Polytechnic University, NY, USA)
John Brooke (University of Manchester, UK)
Martin Buecker (Aachen University, Germany)
Rajkumar Buyya (University of Melbourne, Australia)
Jose Sierra-Camara (University Carlos III of Madrid, Spain)
Shyi-Ming Chen (National Taiwan University of Science and Technology,
Taiwan)
YoungSik Choi (University of Missouri, USA)
Hyunseung Choo (Sungkyunkwan University, Korea)
Bastien Chopard (University of Geneva, Switzerland)
Min Young Chung (Sungkyunkwan University, Korea)
Yiannis Cotronis (University of Athens, Greece)
Danny Crookes (Queen's University of Belfast, UK)
Jose C. Cunha (New University of Lisbon, Portugal)
Brian J. d'Auriol (University of Texas at El Paso, USA)
Alexander Degtyarev (Institute for High Performance Computing
and Data Bases, Russia)
Frederic Desprez (INRIA, France)
Tom Dhaene (University of Antwerp, Belgium)
Beniamino Di Martino (Second University of Naples, Italy)
Hassan Diab (American University of Beirut, Lebanon)
Ivan Dimov (Bulgarian Academy of Sciences, Bulgaria)
Iain Duff (Rutherford Appleton Laboratory, UK and CERFACS, France)
Thom Dunning (NCSA and University of Illinois, USA)
Fabrizio Gagliardi (Microsoft, USA)
Marina L. Gavrilova (University of Calgary, Canada)
Michael Gerndt (Technical University of Munich, Germany)
Osvaldo Gervasi (University of Perugia, Italy)
Bob Gingold (Australian National University, Australia)
James Glimm (SUNY Stony Brook, USA)

Christopher Gold (Hong Kong Polytechnic University, Hong Kong)
Yuriy Gorbachev (Institute of High Performance Computing
and Information Systems, Russia)
Andrzej Goscinski (Deakin University, Australia)
Jin Hai (Huazhong University of Science and Technology, China)
Ladislav Hluchy (Slovak Academy of Science, Slovakia)
Xiaohua Hu (Drexel University, USA)
Eui-Nam John Huh (Seoul Women's University, Korea)
Shen Hong (Japan Advanced Institute of Science and Technology, Japan)
Paul Hovland (Argonne National Laboratory, USA)
Andres Iglesias (University of Cantabria, Spain)
Peter K. Jimack (University of Leeds, UK)
In-Jae Jeong (Hanyang University, Korea)
Chris Johnson (University of Utah, USA)
Benjoe A. Juliano (California State University at Chico, USA)
Peter Kacsuk (MTA SZTAKI Research Institute, Hungary)
Kyung Wo Kang (KAIST, Korea)
Carl Kesselman (USC/ Information Sciences Institute, USA)
Daniel Kidger (Quadrics, UK)
Haeng Kon Kim (Catholic University of Daegu, Korea)
Jin Suk Kim (KAIST, Korea)
Tai-Hoon Kim (Korea Information Security Agency, Korea)
Yoonhee Kim (Syracuse University, USA)
Mike Kirby (University of Utah, USA)
Dieter Kranzmueller (Johannes Kepler University Linz, Austria)
Deok-Soo Kim (Hanyang University, Korea)
Vipin Kumar (University of Minnesota, USA)
Domenico Laforenza (Italian National Research Council, Italy)
Antonio Laganà (University of Perugia, Italy)
Joseph Landman (Scalable Informatics LLC, USA)
Francis Lau (The University of Hong Kong, Hong Kong)
Bong Hwan Lee (Texas A&M University, USA)
Dong Chun Lee (Howon University, Korea)
Dong-Ho Lee (Institute of High Performance Computing, Singapore)
Sang Yoon Lee (Georgia Institute of Technology, USA)
Tae-Jin Lee (Sungkyunkwan University, Korea)
Bogdan Lesyng (ICM Warszawa, Poland)
Zhongze Li (Chinese Academy of Sciences, China)
Laurence Liew (Scalable Systems Pte, Singapore)
David Lombard (Intel Corporation, USA)
Emilio Luque (University Autònoma of Barcelona, Spain)
Michael Mascagni (Florida State University, USA)
Graham Megson (University of Reading, UK)
John G. Michopoulos (US Naval Research Laboratory, USA)
Edward Moreno (Euripides Foundation of Marilia, Brazil)

Youngsong Mun (Soongsil University, Korea)
 Jiri Nedoma (Academy of Sciences of the Czech Republic, Czech Republic)
 Genri Norman (Russian Academy of Sciences, Russia)
 Stephan Olariu (Old Dominion University, USA)
 Salvatore Orlando (University of Venice, Italy)
 Robert Panoff (Shodor Education Foundation, USA)
 Marcin Paprzycki (Oklahoma State University, USA)
 Gyung-Leen Park (University of Texas, USA)
 Ron Perrott (Queen's University of Belfast, UK)
 Dimitri Plemenos (University of Limoges, France)
 Richard Ramaroson (ONERA, France)
 Rosemary Renaut (Arizona State University, USA)
 René S. Renner (California State University at Chico, USA)
 Paul Roe (Queensland University of Technology, Australia)
 Alexey S. Rodionov (Russian Academy of Sciences, Russia)
 Heather J. Ruskin (Dublin City University, Ireland)
 Ole Saastad (Scali, Norway)
 Muhammad Sarfraz (King Fahd University of Petroleum and Minerals,
 Saudi Arabia)
 Edward Seidel (Louisiana State University, USA and Albert-Einstein-Institut,
 Potsdam, Germany)
 Jie Shen (University of Michigan, USA)
 Dale Shires (US Army Research Laboratory, USA)
 Vaclav Skala (University of West Bohemia, Czech Republic)
 Burton Smith (Cray, USA)
 Masha Sosonkina (Ames Laboratory, USA)
 Alexei Sourin (Nanyang Technological University, Singapore)
 Elena Stankova (Institute for High Performance Computing and Data Bases,
 Russia)
 Gunther Stuer (University of Antwerp, Belgium)
 Kokichi Sugihara (University of Tokyo, Japan)
 Boleslaw Szymanski (Rensselaer Polytechnic Institute, USA)
 Ryszard Tadeusiewicz (AGH University of Science and Technology, Poland)
 C.J. Kenneth Tan (OptimaNumerics, UK and Queen's University
 of Belfast, UK)
 David Taniar (Monash University, Australia)
 John Taylor (Streamline Computing, UK)
 Ruppa K. Thulasiram (University of Manitoba, Canada)
 Pavel Tvrdik (Czech Technical University, Czech Republic)
 Putchong Uthayopas (Kasetsart University, Thailand)
 Mario Valle (Swiss National Supercomputing Centre, Switzerland)
 Marco Vanneschi (University of Pisa, Italy)
 Piero Giorgio Verdini (University of Pisa and Istituto Nazionale di Fisica
 Nucleare, Italy)
 Jesus Vigo-Aguar (University of Salamanca, Spain)

Jens Volkert (University of Linz, Austria)
Koichi Wada (University of Tsukuba, Japan)
Stephen Wismath (University of Lethbridge, Canada)
Kevin Wadleigh (Hewlett Packard, USA)
Jerzy Wasniewski (Technical University of Denmark, Denmark)
Paul Watson (University of Newcastle Upon Tyne, UK)
Jan Weglarz (Poznan University of Technology, Poland)
Tim Wilkens (Advanced Micro Devices, USA)
Roman Wyrzykowski (Technical University of Czestochowa, Poland)
Jinchao Xu (Pennsylvania State University, USA)
Chee Yap (New York University, USA)
Osman Yasar (SUNY at Brockport, USA)
George Yee (National Research Council and Carleton University, Canada)
Yong Xue (Chinese Academy of Sciences, China)
Igor Zacharov (SGI Europe, Switzerland)
Xiaodong Zhang (College of William and Mary, USA)
Aledander Zhmakin (SoftImpact, Russia)
Krzysztof Zielinski (ICS UST / CYFRONET, Poland)
Albert Zomaya (University of Sydney, Australia)

Sponsoring Organizations

Institute of Electrical Engineers (IEE), UK
University of Perugia, Italy
University of Calgary, Canada
University of Minnesota, USA
Queen's University of Belfast, UK
The European Research Consortium for Informatics and Mathematics (ERCIM)
The 6th European Framework Project "Distributed European Infrastructure
for Supercomputing Applications" (DEISA)
OptimaNumerics, UK
INTEL
AMD

Table of Contents – Part III

Workshop on Approaches or Methods of Security Engineering (AMSE 2006, Sess. A)

A Security Requirement Management Database Based on ISO/IEC 15408 <i>Shoichi Morimoto, Daisuke Horie, Jingde Cheng</i>	1
Development of Committee Neural Network for Computer Access Security System <i>A. Sermet Anagun</i>	11
C-TOBI-Based Pitch Accent Prediction Using Maximum-Entropy Model <i>Byeongchang Kim, Gary Geunbae Lee</i>	21
Design and Fabrication of Security and Home Automation System <i>Eung Soo Kim, Min Sung Kim</i>	31
PGNIDS(Pattern-Graph Based Network Intrusion Detection System) Design <i>Byung-kwan Lee, Seung-hae Yang, Dong-Hyuck Kwon, Dai-Youn Kim</i>	38
Experiments and Hardware Countermeasures on Power Analysis Attacks <i>ManKi Ahn, HoonJae Lee</i>	48
Information System Modeling for Analysis of Propagation Effects and Levels of Damage <i>InJung Kim, YoonJung Chung, YoungGyo Lee, Eul Gyu Im, Dongho Won</i>	54
A Belt-Zone Method for Decreasing Control Messages in Ad Hoc Networks <i>Youngrag Kim, JaeYoun Jung, Seunghwan Lee, Chonggun Kim</i>	64
A VLSM Address Management Method for Variable IP Subnetting <i>SeongKwon Cheon, DongXue Jin, ChongGun Kim</i>	73
SDSEM: Software Development Success Evolution Model <i>Haeng-Kon Kim, Sang-Yong Byun</i>	84

A Robust Routing Protocol by a Substitute Local Path in Ad Hoc Networks <i>Mary Wu, SangJoon Jung, Seunghwan Lee, Chonggun Kim</i>	93
Power Efficient Wireless LAN Using 16-State Trellis-Coded Modulation for Infrared Communications <i>Hae Geun Kim</i>	104
The Design and Implementation of Real-Time Environment Monitoring Systems Based on Wireless Sensor Networks <i>Kyung-Hoon Jung, Seok-Cheol Lee, Hyun-Suk Hwang, Chang-Soo Kim</i>	115
Ontology-Based Information Search in the Real World Using Web Services <i>Hyun-Suk Hwang, Kyoo-Seok Park, Chang-Soo Kim</i>	125
An Active Node Set Maintenance Scheme for Distributed Sensor Networks <i>Tae-Young Byun, Minsu Kim, Sungho Hwang, Sung-Eok Jeon</i>	134
Intelligent Information Search Mechanism Using Filtering and NFC Based on Multi-agents in the Distributed Environment <i>Subong Yi, Bobby D. Gerardo, Young-Seok Lee, Jaewan Lee</i>	144
Network Anomaly Behavior Detection Using an Adaptive Multiplex Detector <i>Misun Kim, Minsoo Kim, JaeHyun Seo</i>	154
Applying Product Line to the Embedded Systems <i>Haeng-Kon Kim</i>	163
Enhanced Fuzzy Single Layer Learning Algorithm Using Automatic Tuning of Threshold <i>Kwang-Baek Kim, Byung-Kwan Lee, Soon-Ho Kim</i>	172
Optimization of Location Management in the Distributed Location-Based Services Using Collaborative Agents <i>Romeo Mark A. Mateo, Jaewan Lee, Hyunho Yang</i>	178
Design of H.264/AVC-Based Software Decoder for Mobile Phone <i>Hyung-Su Jeon, Hye-Min Noh, Cheol-Jung Yoo, Ok-Bae Chang</i>	188
Transforming a Legacy System into Components <i>Haeng-Kon Kim, Youn-Ky Chung</i>	198

Pseudorandom Number Generator Using Optimal Normal Basis <i>Injoo Jang, Hyeong Seon Yoo</i>	206
Efficient Nonce-Based Authentication Scheme Using Token-Update <i>Wenbo Shi, Hyeong Seon Yoo</i>	213
An Efficient Management of Network Traffic Performance Using Framework-Based Performance Management Tool <i>Seong-Man Choi, Cheol-Jung Yoo, Ok-Bae Chang</i>	222
A Prediction Method of Network Traffic Using Time Series Models <i>Sangjoon Jung, Chonggun Kim, Younky Chung</i>	234
An Obstacle Avoidance Method for Chaotic Robots Using Angular Degree Limitations <i>Youngchul Bae, MalRey Lee, Thomas M. Gatton</i>	244
Intersection Simulation System Based on Traffic Flow Control Framework <i>Chang-Sun Shin, Dong-In Ahn, Hyun Yoe, Su-Chong Joo</i>	251
A HIICA(Highly-Improved Intra CA) Design for M-Commerce <i>Byung-kwan Lee, Chang-min Kim, Dae-won Shin, Seung-hae Yang</i>	261
Highly Reliable Synchronous Stream Cipher System for Link Encryption <i>HoonJae Lee</i>	269
Recognition of Concrete Surface Cracks Using ART2-Based Radial Basis Function Neural Network <i>Kwang-Baek Kim, Hwang-Kyu Yang, Sang-Ho Ahn</i>	279
Hybrid Image Mosaic Construction Using the Hierarchical Method <i>Oh-Hyung Kang, Ji-Hyun Lee, Yang-Won Rhee</i>	287
Workshop on Applied Cryptography and Information Security (ACIS 2006)	
Public Key Encryption with Keyword Search Based on K-Resilient IBE <i>Dalia Khader</i>	298
A Generic Construction of Secure Signatures Without Random Oracles <i>Jin Li, Yuen-Yan Chan, Yanming Wang</i>	309

A Separation Between Selective and Full-Identity Security Notions for Identity-Based Encryption <i>David Galindo</i>	318
Traceable Signature: Better Efficiency and Beyond <i>He Ge, Stephen R. Tate</i>	327
On the TYS Signature Scheme <i>Marc Joye, Hung-Mei Lin</i>	338
Efficient Partially Blind Signatures with Provable Security <i>Qianhong Wu, Willy Susilo, Yi Mu, Fanguo Zhang</i>	345
A Framework for Robust Group Key Agreement <i>Jens-Matthias Bohli</i>	355
BGN Authentication and Its Extension to Convey Message Commitments <i>Yuen-Yan Chan, Jin Li</i>	365
New Security Problem in RFID Systems “Tag Killing” <i>Dong-Guk Han, Tsuyoshi Takagi, Ho Won Kim, Kyo Il Chung</i>	375
A Model for Security Vulnerability Pattern <i>Hyungwoo Kang, Kibom Kim, Soonjwa Hong, Dong Hoon Lee</i>	385
A New Timestamping Scheme Based on Skip Lists <i>Kaouthar Blibech, Alban Gabillon</i>	395
A Semi-fragile Watermarking Scheme Based on SVD and VQ Techniques <i>Hsien-Chu Wu, Chuan-Po Yeh, Chwei-Shyong Tsai</i>	406
New Constructions of Universal Hash Functions Based on Function Sums <i>Khoongming Khoo, Swee-Huay Heng</i>	416
Analysis of Fast Blockcipher-Based Hash Functions <i>Martin Stanek</i>	426
Application of LFSRs for Parallel Sequence Generation in Cryptologic Algorithms <i>Sourav Mukhopadhyay, Palash Sarkar</i>	436
Provable Security for an RC6-like Structure and a MISTY-FO-like Structure Against Differential Cryptanalysis <i>Changhoon Lee, Jongsung Kim, Jaechul Sung, Seokhie Hong, Sangjin Lee</i>	446

Design and Implementation of an FPGA-Based 1.452-Gbps Non-pipelined AES Architecture <i>Ignacio Algreto-Badillo, Claudia Feregrino-Uribe, René Cumplido . . .</i>	456
---	-----

Workshop on Internet Communications Security (WICS 2006)

Security Weaknesses in Two Proxy Signature Schemes <i>Jiqiang Lu</i>	466
A Proposal of Extension of FMS-Based Mechanism to Find Attack Paths <i>Byung-Ryong Kim, Ki-Chang Kim</i>	476
Comparative Analysis of IPv6 VPN Transition in NEMO Environments <i>Hyung-Jin Lim, Dong-Young Lee, Tai-Myoung Chung</i>	486
A Short-Lived Key Selection Approach to Authenticate Data Origin of Multimedia Stream <i>Namhi Kang, Younghan Kim</i>	497
Weakest Link Attack on Single Sign-On and Its Case in SAML V2.0 Web SSO <i>Yuen-Yan Chan</i>	507
An Inter-domain Key Agreement Protocol Using Weak Passwords <i>Youngsook Lee, Junghyun Nam, Dongho Won</i>	517
A Practical Solution for Distribution Rights Protection in Multicast Environments <i>Josep Pegueroles, Marcel Fernández, Francisco Rico-Novella, Miguel Soriano</i>	527
Audit-Based Access Control in Nomadic Wireless Environments <i>Francesco Palmieri, Ugo Fiore</i>	537

Workshop on Optimization: Theories and Applications (OTA 2006)

Cost – Time Trade Off Models Application to Crashing Flow Shop Scheduling Problems <i>Morteza Bagherpour, Siamak Noori, S. Jafar Sadjadi</i>	546
--	-----

The ASALB Problem with Processing Alternatives Involving Different Tasks: Definition, Formalization and Resolution <i>Liliana Capacho, Rafael Pastor</i>	554
Satisfying Constraints for Locating Export Containers in Port Container Terminals <i>Kap Hwan Kim, Jong-Sool Lee</i>	564
A Price Discrimination Modeling Using Geometric Programming <i>Seyed J. Sadjadi, M. Ziaee</i>	574
Hybrid Evolutionary Algorithms for the Rectilinear Steiner Tree Problem Using Fitness Estimation <i>Byounghak Yang</i>	581
Data Reduction for Instance-Based Learning Using Entropy-Based Partitioning <i>Seung-Hyun Son, Jae-Yearn Kim</i>	590
Coordinated Inventory Models with Compensation Policy in a Three Level Supply Chain <i>Jeong Hun Lee, Il Kyeong Moon</i>	600
Using Constraint Satisfaction Approach to Solve the Capacity Allocation Problem for Photolithography Area <i>Shu-Hsing Chung, Chun-Ying Huang, Amy Hsin-I Lee</i>	610
Scheduling an R&D Project with Quality-Dependent Time Slots <i>Mario Vanhoucke</i>	621
The Bottleneck Tree Alignment Problems <i>Yen Hung Chen, Chuan Yi Tang</i>	631
Performance Study of a Genetic Algorithm for Sequencing in Mixed Model Non-permutation Flowshops Using Constrained Buffers <i>Gerrit Färber, Anna M. Coves Moreno</i>	638
Optimizing Relative Weights of Alternatives with Fuzzy Comparative Judgment <i>Chung-Hsing Yeh, Yu-Hern Chang</i>	649
Model and Solution for the Multilevel Production-Inventory System Before Ironmaking in Shanghai Baoshan Iron and Steel Complex <i>Guoli Liu, Lixin Tang</i>	659

A Coordination Algorithm for Deciding Order-Up-To Level of a Serial Supply Chain in an Uncertain Environment <i>Kung-Jeng Wang, Wen-Hai Chih, Ken Hwang</i>	668
Optimization of Performance of Genetic Algorithm for 0-1 Knapsack Problems Using Taguchi Method <i>A.S. Anagun, T. Sarac</i>	678
Truck Dock Assignment Problem with Time Windows and Capacity Constraint in Transshipment Network Through Crossdocks <i>Andrew Lim, Hong Ma, Zhaowei Miao</i>	688
An Entropy Based Group Setup Strategy for PCB Assembly <i>In-Jae Jeong</i>	698
Cross-Facility Production and Transportation Planning Problem with Perishable Inventory <i>Sandra Duni Ekşioğlu, Mingzhou Jin</i>	708
A Unified Framework for the Analysis of M/G/1 Queue Controlled by Workload <i>Ho Woo Lee, Se Won Lee, Won Ju Seo, Sahng Hoon Cheon, Jongwoo Jeon</i>	718
Tabu Search Heuristics for Parallel Machine Scheduling with Sequence-Dependent Setup and Ready Times <i>Sang-Il Kim, Hyun-Seon Choi, Dong-Ho Lee</i>	728
The Maximum Integer Multiterminal Flow Problem <i>Cédric Bentz</i>	738
Routing with Early Ordering for Just-In-Time Manufacturing Systems <i>Mingzhou Jin, Kai Liu, Burak Eksioğlu</i>	748
A Variant of the Constant Step Rule for Approximate Subgradient Methods over Nonlinear Networks <i>Eugenio Mijangos</i>	757
On the Optimal Buffer Allocation of an FMS with Finite In-Process Buffers <i>Soo-Tae Kwon</i>	767
Optimization Problems in the Simulation of Multifactor Portfolio Credit Risk <i>Wanmo Kang, Kyungsik Lee</i>	777

Two-Server Network Disconnection Problem <i>Byung-Cheon Choi, Sung-Pil Hong</i>	785
One-Sided Monge TSP Is NP-Hard <i>Vladimir Deineko, Alexander Tiskin</i>	793
On Direct Methods for Lexicographic Min-Max Optimization <i>Włodzimierz Ogryczak, Tomasz Śliwiński</i>	802
Multivariate Convex Approximation and Least-Norm Convex Data-Smoothing <i>Alex Y.D. Siem, Dick den Hertog, Aswin L. Hoffmann</i>	812
Linear Convergence of Tatônnement in a Bertrand Oligopoly <i>Guillermo Gallego, Woonghee Tim Huh, Wanmo Kang, Robert Phillips</i>	822
Design for Using Purpose of Assembly-Group <i>Hak-Soo Mok, Chang-Hyo Han, Chan-Hyoung Lim, John-Hee Hong, Jong-Rae Cho</i>	832
A Conditional Gaussian Martingale Algorithm for Global Optimization <i>Manuel L. Esquivel</i>	841
Finding the Number of Clusters Minimizing Energy Consumption of Wireless Sensor Networks <i>Hyunsoo Kim, Hee Yong Youn</i>	852
A Two-Echelon Deteriorating Production-Inventory Newsboy Model with Imperfect Production Process <i>Hui-Ming Wee, Chun-Jen Chung</i>	862
Mathematical Modeling and Tabu Search Heuristic for the Traveling Tournament Problem <i>Jin Ho Lee, Young Hoon Lee, Yun Ho Lee</i>	875
An Integrated Production-Inventory Model for Deteriorating Items with Imperfect Quality and Shortage Backordering Considerations <i>H.M. Wee, Jonas C.P. Yu, K.J. Wang</i>	885
A Clustering Algorithm Using the Ordered Weight Sum of Self-Organizing Feature Maps <i>Jong-Sub Lee, Maing-Kyu Kang</i>	898

Global Optimization of the Scenario Generation and Portfolio Selection Problems	
<i>Panos Parpas, Berç Rustem</i>	908
A Generalized Fuzzy Optimization Framework for R&D Project Selection Using Real Options Valuation	
<i>E. Ertugrul Karsak</i>	918
Supply Chain Network Design and Transshipment Hub Location for Third Party Logistics Providers	
<i>Seungwoo Kwon, Kyungdo Park, Chulung Lee, Sung-Shick Kim, Hak-Jin Kim, Zhong Liang</i>	928
A Group Search Optimizer for Neural Network Training	
<i>S. He, Q.H. Wu, J.R. Saunders</i>	934
Application of Two-Stage Stochastic Linear Program for Portfolio Selection Problem	
<i>Kuo-Hwa Chang, Huifen Chen, Ching-Fen Lin</i>	944
General Tracks	
Hierarchical Clustering Algorithm Based on Mobility in Mobile Ad Hoc Networks	
<i>Sulyun Sung, Yuhwa Seo, Yongtae Shin</i>	954
An Alternative Approach to the Standard Enterprise Resource Planning Life Cycle: Enterprise Reference Metamodeling	
<i>Miguel Gutiérrez, Alfonso Durán, Pedro Cocho</i>	964
Static Analysis Based Software Architecture Recovery	
<i>Jiang Guo, Yuehong Liao, Raj Pamula</i>	974
A First Approach to a Data Quality Model for Web Portals	
<i>Angelica Caro, Coral Calero, Ismael Caballero, Mario Piattini</i>	984
Design for Environment-Friendly Product	
<i>Hak-Soo Mok, Jong-Rae Cho, Kwang-Sup Moon</i>	994
Performance of HECC Coprocessors Using Inversion-Free Formulae	
<i>Thomas Wollinger, Guido Bertoni, Luca Breveglieri, Christof Paar</i>	1004
Metrics of Password Management Policy	
<i>Carlos Villarrubia, Eduardo Fernández-Medina, Mario Piattini</i>	1013

Using UML Packages for Designing Secure Data Warehouses <i>Rodolfo Villarroel, Emilio Soler, Eduardo Fernández-Medina, Juan Trujillo, Mario Piattini</i>	1024
Practical Attack on the Shrinking Generator <i>Pino Caballero-Gil, Amparo Fúster-Sabater</i>	1035
A Comparative Study of Proposals for Establishing Security Requirements for the Development of Secure Information Systems <i>Daniel Mellado, Eduardo Fernández-Medina, Mario Piattini</i>	1044
Stochastic Simulation Method for the Term Structure Models with Jump <i>Kisoeb Park, Moonseong Kim, Seki Kim</i>	1054
The Ellipsoidal l_p Norm Obnoxious Facility Location Problem <i>Yu Xia</i>	1064
On the Performance of Recovery Rate Modeling <i>J. Samuel Baixauli, Susana Alvarez</i>	1073
Using Performance Profiles to Evaluate Preconditioners for Iterative Methods <i>Michael Lazzareschi, Tzu-Yi Chen</i>	1081
Multicast ω -Trees Based on Statistical Analysis <i>Moonseong Kim, Young-Cheol Bang, Hyunseung Choo</i>	1090
The Gateways Location and Topology Assignment Problem in Hierarchical Wide Area Networks: Algorithms and Computational Results <i>Przemyslaw Ryba, Andrzej Kasprzak</i>	1100
Developing an Intelligent Supplier Chain System Collaborating with Customer Relationship Management <i>Gye Hang Hong, Sung Ho Ha</i>	1110
The Three-Criteria Servers Replication and Topology Assignment Problem in Wide Area Networks <i>Marcin Markowski, Andrzej Kasprzak</i>	1119
An Efficient Multicast Tree with Delay and Delay Variation Constraints <i>Moonseong Kim, Young-Cheol Bang, Jong S. Yang, Hyunseung Choo</i>	1129
Algorithms on Extended (δ, γ) -Matching <i>Inbok Lee, Raphaël Clifford, Sung-Ryul Kim</i>	1137

SOM and Neural Gas as Graduated Nonconvexity Algorithms <i>Ana I. González, Alicia D’Anjou, M. Teresa García-Sebastian, Manuel Graña</i>	1143
Analysis of Multi-domain Complex Simulation Studies <i>James R. Gattiker, Earl Lawrence, David Higdon</i>	1153
A Fast Method for Detecting Moving Vehicles Using Plane Constraint of Geometric Invariance <i>Dong-Joong Kang, Jong-Eun Ha, Tae-Jung Lho</i>	1163
Robust Fault Matched Optical Flow Detection Using 2D Histogram <i>Jaechoon Chon, Hyongsuk Kim</i>	1172
Iris Recognition: Localization, Segmentation and Feature Extraction Based on Gabor Transform <i>Mohammadreza Noruzi, Mansour Vafadoost, M. Shahram Moin</i>	1180
Optimal Edge Detection Using Perfect Sharpening of Ramp Edges <i>Eun Mi Kim, Cheryl Soo Park, Jong Gu Lee</i>	1190
Eye Tracking Using Neural Network and Mean-Shift <i>Eun Yi Kim, Sin Kuk Kang</i>	1200
The Optimal Feature Extraction Procedure for Statistical Pattern Recognition <i>Marek Kurzynski, Edward Puchala</i>	1210
A New Approach for Human Identification Using Gait Recognition <i>Murat Ekinçi</i>	1216
Author Index	1227

A Security Requirement Management Database Based on ISO/IEC 15408

Shoichi Morimoto¹, Daisuke Horie², and Jingde Cheng²

¹ Advanced Institute of Industrial Technology,
1-10-40, Higashi-ōi, Shinagawa-ku, Tokyo, 140-0011, Japan
morimo@aise.ics.saitama-u.ac.jp

² Department of Information and Computer Sciences, Saitama University,
Saitama, 338-8570, Japan
{morimo, horie, cheng}@aise.ics.saitama-u.ac.jp

Abstract. With the scale-spreading and diversification of information systems, security requirements for the systems are being more and more complicated. It is desirable to apply database technologies to information security engineering in order to manage the security requirements in design and development of the systems. This paper proposes a security requirement management database based on the international standard ISO/IEC 15408 that defines security functional requirements which should be satisfied by various information systems. The database can aid design and development of information systems that require high security such that it enables to suitably refer to required data of security requirements.

1 Introduction

Nowadays, in design and development of various information systems, it is necessary to take security issues into consideration and to verify whether or not the systems satisfy stringent security criteria. Thus, ISO/IEC 15408 was established as a criterion for evaluating the security level of IT products and information systems [7]. ISO/IEC 15408 defines security functional requirements which should be applied to validate an information system. Developers have to make a security design document for evaluation of ISO/IEC 15408. The process of making the documents is very complicate. Moreover, they must also decide by themselves which security functional requirements are necessary to their systems. As it is, it is difficult to determine which requirements are required. Thus, it is a very hard task and difficult to develop information systems which comply with ISO/IEC 15408.

On the other hand, in software engineering, especially in requirement engineering, some databases have been proposed in order to collect, manage and reuse the past knowledge/experience in information system design and development, e.g., [8, 6, 13]. Database technologies have successfully been applied to software engineering. Similarly, one can manage the knowledge/experience for security requirements in information system design and development that comply with ISO/IEC 15408 by a database.

This paper proposes a security requirement management database based on the international standard ISO/IEC 15408, named “ISEDS (Information Security Engineering Database System).” Users of ISEDS can collect, manage and reuse security requirements. Thus, ISEDS can aid design and development of secure information systems, which satisfy the security criteria of ISO/IEC 15408.

2 ISO/IEC 15408

We herein explain ISO/IEC 15408 which is the base of ISEDS. ISO/IEC 15408 consists of three parts, i.e., “Part 1: Introduction and general model,” “Part 2: Security functional requirements,” and “Part 3: Security assurance requirements.” Part 1 is the introduction of ISO/IEC 15408. Part 1 provides that sets of documents, so-called ‘security targets,’ must be created and submitted for evaluating information systems by ISO/IEC 15408. In order to simplify the creation of security targets, Part 1 also proposes templates called ‘protection profiles.’ Part 2 establishes a set of functional components as a standard way of expressing the functional requirements for target information systems. In other words, Part 2 defines the requirements for security functions which should be applied to validate an information system. Part 3 establishes a set of assurance components as a standard way of expressing the assurance requirements for target information systems.

We defined the structure of ISEDS according to the structure of the documents and Part 2.

2.1 The Documents for Evaluation of ISO/IEC 15408

Applicants who apply to obtain the evaluation of ISO/IEC 15408 have to describe and submit a security target to the evaluation organization. A security target, ST for short, must describe range of a target information system which is evaluated (target of evaluation, TOE for short), assumed threats in TOE, security objectives to oppose these threats, functions required for achievement of these objectives, cf., Fig. 1 on the next page. In particular, the required security functions must be quoted from security functional requirements of Part 2.

A protection profile, PP for short, defines a set of implementation independent IT security requirements for a category of information systems, e.g., OS, DBMS, Firewall, etc. A PP consists, roughly speaking, of threats which may be assumed in the PP’s category, security objectives that oppose the assumed threats, and functions required in order to achieve the security objectives. An ST is then created by instantiating a PP. In the documents complying with ISO/IEC 15408, security requirements are analyzed and defined in the above procedure.

2.2 The Structure of Security Functional Requirements

Part 2 has a hierarchical structure which is composed by classes, families, components, and elements in the order (Fig. 2).

Each element is an indivisible security requirement. The components consist of the smallest selectable set of the elements that may be included in specifications. Furthermore, the components may be mutually dependent. The families

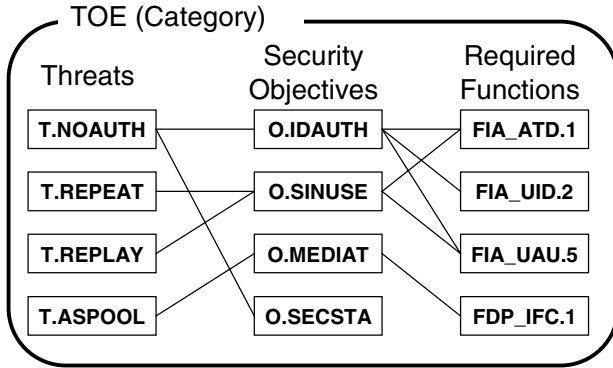


Fig. 1. The document structure of STs or PPs

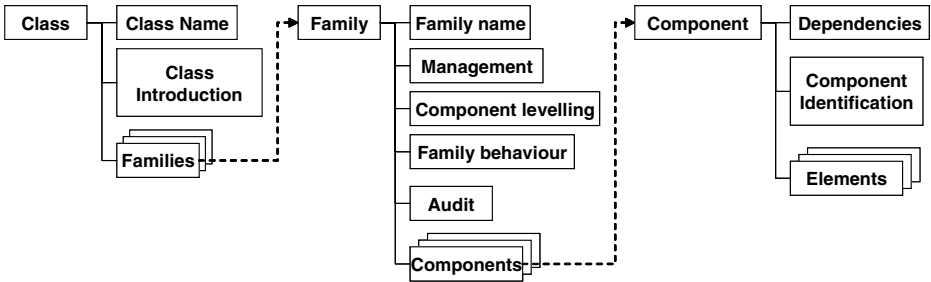


Fig. 2. The hierarchical structure of ISO/IEC 15408 Part2

are a group of the components that share security objectives but may differ in emphasis or rigor. The classes are a group of the families that share common focuses. Security functional requirements exactly and directly are described in the elements of the bottom layer. The following text is one of the security functional requirements.

Class FCO: Communication

This class provides ... (omitted)

FCO_NRO Non-repudiation of origin

Family behavior

...

Management: FCO_NRO.1, FCO_NRO.2

...

Audit: FCO_NRO.1

...

FCO_NRO.1 Selective proof of origin

Hierarchical to: No other components.

FCO_NRO.1.1 The TSF shall be able to generate evidence of origin for transmitted [assignment: list of information types] at the request of the [selection: originator, recipient, [assignment: list of third parties]].

...

Dependencies: FIA_UID.1 Timing of identification

This text shows the element FCO_NRO.1.1, the component FCO_NRO.1, the family FCO_NRO, and the class FCO. The element directly describes the requirement in natural language.

We adopted the structure of the documents and the security functional requirements as the structure of ISEDS. Since it is suitable for expressing the above structures, we developed ISEDS as a relational database.

3 Design and Implementation of ISEDS

We designed ISEDS based on the structures mentioned above. The following is the detail of the design and the implementation.

3.1 The Schema Design

We defined the structure of ISEDS as a set of a category, threats, objectives, and functions. We show some examples of STs or PPs in order to clarify the structure. The original text delineating one of the threats in a traffic-filter firewall PP is as follows [5].

T.NOAUTH Illegal access

An unauthorized person may attempt to bypass the security of the TOE so as to access and use security functions and/or non-security functions provided by the TOE.

T.NOAUTH specifies that unauthorized persons may attack the system. One of the security objectives which resist this threat is described as follows.

O.IDAUTH Authentication

The TOE must uniquely identify and authenticate the claimed identity of all users, before granting a user access to TOE functions. This security objective is necessary to counter the threat: T.NOAUTH because it requires that users be uniquely identified before accessing the TOE.

The other objectives are described as O.SECSTA, O.ENCRYPT, O.SELPRO, O.SECFUN, and O.LIMEXT. Contrary, one security objective may resist many threats. Thus, the relationship of threats and objectives is M:M (many to many).

In order to achieve O.IDAUTH, the PP describes that the elements FIA_AT-D.1.1, FIA_UID.2.1, FIA_UAU.5.1, and FIA_UAU.5.2 are required. On the other hand, one element may be required for achievement of many objectives sometimes. Thus, the relationship of objectives and elements is M:M as well as threats and objectives.

For implementation of the required elements, an ST must describe that they are actually implemented as what functions in information systems. These functions are called TOE security functions, TSF for short. The following is a certain TSF in an ST of PKI software for smart cards [9].

SF.PINLENGTHMANAGE**SF1**

Issuer can set Minimum length of administrator/normal user PIN before issuing the MULTOS smart card. After the MULTOS smart card is issued these values cannot be changed.

SF2

When administrator/normal user tries to change one's PIN and inputs new PIN shorter than Minimum length of administrator/normal user PIN, the TOE denies the change of PIN.

First, an ST describes high level TSFs, e.g., SF.PINLENGTHMANAGE. Next, it details a high level TSF as low level TSFs with top-down, e.g., SF1 and SF2. That is, the relationship of high level TSFs and low level TSFs is 1:M (one to many). Moreover, one element matches many high level TSFs. Contrary, one high level TSF may implement many elements. Thus, the relationship of them is M:M.

As mentioned above, it turns out that entities of ISEDS are classified as follows.

(a) TOEs, **(b)** Threats, **(c)** Security objectives, **(d)** High level TSFs, **(e)** Low level TSFs, **(f)** Classes, **(g)** Families, **(h)** Components, **(i)** Elements

The cardinality of these entities also becomes clear in Table. 1.

Table 1. The cardinality of the entities

	a	b	c	d	e	f	g	h	i
a	-	1:M	1:M	1:M	-	-	-	-	-
b	1:M	-	M:M	-	-	-	-	-	-
c	1:M	M:M	-	-	-	-	-	-	M:M
d	1:M	-	-	-	1:M	-	-	-	-
e	-	-	-	1:M	-	-	-	-	M:M
f	-	-	-	-	-	-	1:M	-	-
g	-	-	-	-	-	1:M	-	1:M	-
h	-	-	-	-	-	-	1:M	M:M	1:M
i	-	-	M:M	-	M:M	-	-	1:M	-

We designed these entities as schemata in a database model based on Fig. 2 and Table. 1 (cf., Fig. 3).

Fig. 3 was drawn with Microsoft Visio Professional 2002. Visio can clearly and easily design accurate database model diagrams in IDEF1X and relational notation [2]. Visio can also automatically generate SQL sentences from the database model diagrams. In Fig. 3, PK denotes primal key and FK denotes foreign key

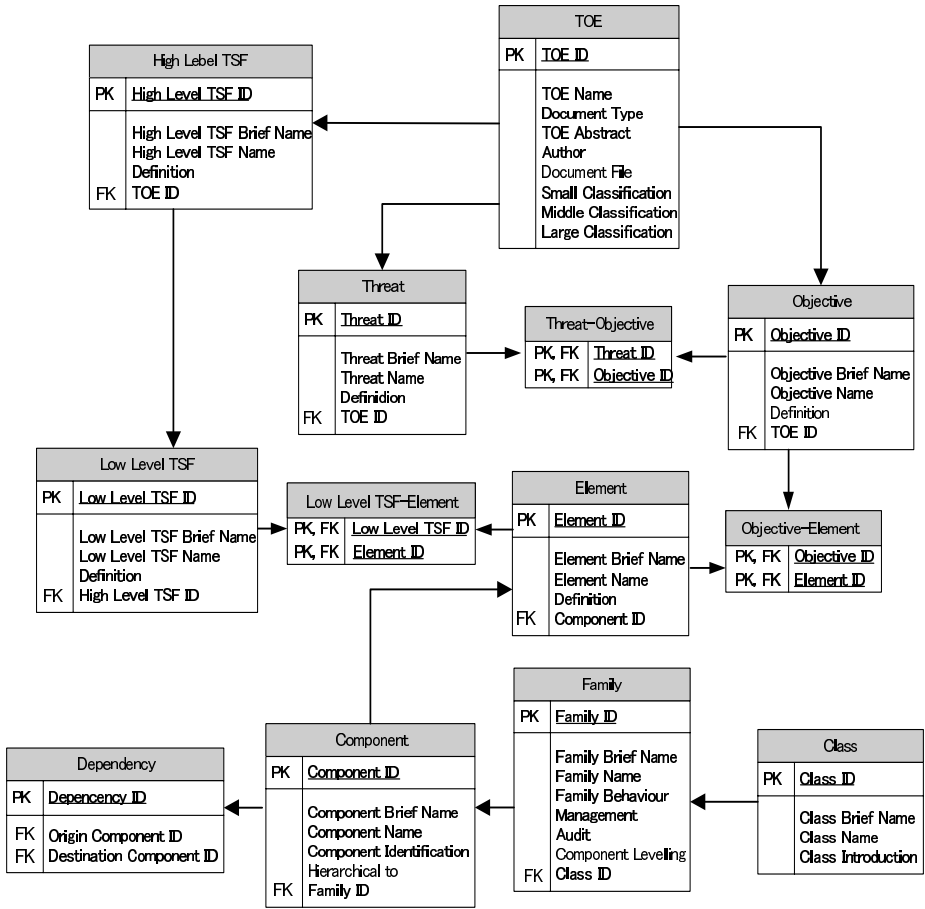


Fig. 3. The database model diagram

in a schema. Attributes in the bold font mean indispensable items of a schema. An arrow denotes a relationship between schemata. The tip of an arrow shows the cardinality M and the starting point of the arrow shows the cardinality 1.

The schema *TOE* has attributes that are written to an ST or a PP, i.e., *TOE Name*, *TOE Abstract*, *Author*, *Small Classification*, *Middle Classification*, and *Large Classification*. The classifications denote TOE kinds. So far, *Large Classification* is only ‘IT products’ now. *Middle Classification* is ‘software,’ ‘hardware,’ ‘middleware’ etc. *Small Classification* shows the concrete TOE kind, e.g., Database, Firewall, IC card, OS, Copier, and so on. In addition to these attributes, we defined *Document Type* and *Document File*. *Document Type* is a flag for distinguishing an ST or a PP. *Document File* is an attribute for storing the binary file of the ST or PP. *TOE* relates to one or more *Threats*, *Objectives*, and *High Level TSFs*.

The schemata *Threat*, *Objective*, and *High Level TSF* have their ID numbers, abbreviation names (e.g., T.NOAUTH, O.IDAUTH, or SF.PINLENGTHMANAGE), formal names (e.g., Illegal access or Authentication), texts of the definition on the documents, and foreign keys to *TOE*. The schema *Low Level TSF* is almost the same as these schemata. The schemata of security functional requirements have the attributes shown in Fig. 2. The dependencies of components are expressed as a schema, because they are the M:M self references.

3.2 The Implementation

We implemented ISEDS based on the design with PostgreSQL 8.1, because PostgreSQL is one of the eminent open source databases and can use virtual tables [12]. We stored the data of all security functional requirements in ISO/IEC 15408 Part 2 into ISEDS. Users of ISEDS can easily retrieve their required parts of Part 2. It is not necessary to turn the document exceeding 350 pages one by one.

Moreover, 653 documents about security targets [3] and 125 documents about protection profiles [4] certified by ISO/IEC 15408 are published on the common criteria portal web site as of November, 2005. We also extracted the data of all of them and stored the extracted data into ISEDS. Therefore, the users can retrieve reliable data, i.e., how threats, objectives, and TSFs were described in the certified information systems. Additionally, the users can also retrieve which security functional requirements were used in the certified information systems.

4 Benefits and Applications

ISEDS can be used as follows.

Users can retrieve what threats, objectives, or functions are required in a category of information systems. For example, it can search what threats are assumed in the firewall system category.

```
SELECT T.Threat_Name, T.Definition FROM TOE, Threat T WHERE
TOE.Small_Classification LIKE '%firewall%' AND TOE.TOE_ID = T.TOE_ID
```

The users can retrieve what category assumes threats and what objectives and functions can resist threats. For example, it can search what objectives resist to spoofing.

```
SELECT O.Objective_Name, O.Definition FROM Objective O, Threat T,
Threat-Objective R WHERE T.Definition LIKE '%spoofing%' AND T.Threat ID
= R.Threat ID AND R.Objective ID = O.Objective ID
```

The users can retrieve what objective is a countermeasure against threats and what functions implement an objective. For example, it can search which elements are required for concealment of IP addresses.


```
SELECT E.Element_Name, E.Definition FROM Objective O, Element E,
Objective-Element R WHERE O.Definition LIKE '%concealment of IP ad-
dress%' AND O.Objective_ID = R.Objective_ID AND R.Element_ID =
E.Element_ID
```

The users can retrieve what categories, threats, and objectives require a security function. For example, it can search what categories require the function of IP packet filtering.

```
SELECT TOE.Small_Classification FROM Low_Level_TSF L, High_Level_TSF
H, TOE WHERE L.Definition LIKE '%IP packet filtering%' AND L.High_Level_
TSF_ID = H.High_Level_TSF_ID AND H.TOE_ID = TOE.TOE_ID
```

Besides these retrievals, the users can retrieve what categories, threats, and objectives require an element. Moreover, various retrievals may be possible by using the hierarchical structure of Part 2. For example, it can search what components are dependent on the components in the family FTA-TSE.

```
SELECT C2.Component_Brief_Name FROM Component C1, Component C2,
Family F, Dependency D WHERE F.Family_Brief_Name = 'FTA-TSE' AND
F.Family_ID = C1.Family_ID AND C1.Component_ID = D.Origin_Component_
ID AND D.Destination_Component_ID = C2.Component_ID
```

Naturally, the users can update ISEDS by defining and storing a new security requirement as a set of a category (TOE), threats, objectives, required elements, and TSFs.

We have already proposed a security specification verification technique based on ISO/IEC 15408 [11]. We beforehand formalized all elements of ISO/IEC 15408 Part 2 as formal criteria [1]. The technique enables strict verification using formal methods and the formal criteria. With the technique, one can strictly verify whether or not information systems designed by ISEDS satisfy the elements of ISO/IEC 15408. Conversely, users of the technique can retrieve elements which are required in the certified information systems similar to a verification target information system. The weakness of the technique is that the users must decide by themselves which elements are necessary to target information systems. ISEDS solves this problem.

Additionally, we have proposed a method which simplifies creation of a security specification in information systems [10]. In the paper, we successfully classified and rearranged PPs in order to make them possible to more efficiently use. The advantage of the method is that even a developer who is relatively inexperienced in security issues can easily create specifications which satisfy security criteria with the rearranged PPs, because PP's security has been guaranteed by ISO/IEC 15408. However, it is hard to rearrange PPs, because it must be carefully considered with reference to many PPs. ISEDS also solves this

problem. Because of this contribution, anyone can easily improve PPs and may solve various security issues of information systems by the improved PPs.

5 Concluding Remarks

In order to apply database technologies to information security engineering, we have designed and developed ISEDS, a security requirement management database based on the international standard ISO/IEC 15408. Users of ISEDS can collect, manage and reuse security requirements for design and development of various information systems in the form according to ISO/IEC 15408. Since we already stored all data of ISO/IEC 15408 Part 2 into ISEDS, the users can also get their required information of Part 2 without reading vast pages of Part 2 document. Thus, ISEDS can mitigate their labor for design and development of secure information systems. ISEDS can also support design and development of information systems which satisfy the security criteria of ISO/IEC 15408. It is verifiable by our verification technique [11]. ISEDS may be applicable to various purposes. We also expect that ISEDS will be a good example for “database-izing” criteria like ISO standard. We are preparing a web site in which users can use ISEDS [1].

A subject of ISEDS is consistency of the data. Definition of threats, security objectives, and security functions in STs or PPs is different for every document case by case. Moreover, some information systems do not have clear classification. Because of the problem, retrieval of required data may be difficult. Thus, we need to define further common format for such data.

Furthermore, the structure of STs, PPs, security requirements, and the security functional requirements can be easily, exactly and rigorously expressed in XML. Therefore, we are improving ISEDS as a native XML database into which users can directly store the XML documents and are developing its web service.

References

1. Advanced Information Systems Engineering Laboratory, Saitama University.: ISEDS: Information Security Engineering Database System. <http://www.aise.ics.saitama-u.ac.jp/>
2. Bruce, T.A.: Designing Quality Databases with IDEF1X Information Models. Dorset House Publishing Company (1991)
3. Common Criteria Portal Org.: Evaluated product files. <http://www.commoncriteriaportal.org/public/files/epfiles/>
4. Common Criteria Portal Org.: Protection profile files. <http://www.commoncriteriaportal.org/public/files/ppfiles/>
5. Dolan, K., Wright, P., Montequin, R., Mayer, B., Gilmore, L., and Hall, C.: U.S. Department of Defense Traffic-Filter Firewall Protection Profile for Medium Robustness Environments. National Security Agency (2001)
6. International Software Benchmarking Standard Group.: Empirical Databases of Metrics Collected from Software Projects. <http://www.isbsg.org/>

7. ISO/IEC 15408 standard.: Information Technology - Security Techniques - Evaluation Criteria for IT Security (1999)
8. Jiao, J. and Tseng, M.: A Requirement Management Database System for Product Definition. *Journal of Integrated Manufacturing Systems*, Vol. 10, No. 3, pp. 146-154 (1999)
9. Miyazawa, T. and Sugawara, H.: Smart Folder 3 Security Target Version: 2.19. Hitachi Software Engineering Co., Ltd., January (2004)
10. Morimoto, S. and Cheng, J.: Patterning Protection Profiles by UML for Security Specifications. *Proceedings of the IEEE 2005 International Conference on Intelligent Agents, Web Technology and Internet Commerce (IAWTIC'05)*, Vol. II, pp. 946-951, Vienna, Austria, November (2005)
11. Morimoto, S., Shigematsu, S., Goto, Y., and Cheng, J.: A Security Specification Verification Technique Based on the International Standard ISO/IEC 15408. *Proceedings of the 21st Annual ACM Symposium on Applied Computing (SAC'06)*, Dijon, France, April (2006)
12. PostgreSQL Global Development Group.: PostgreSQL.
<http://www.postgresql.org/>
13. Software Engineering Institute.: Software Engineering Information Repository.
<http://seir.sei.cmu.edu/>

Development of Committee Neural Network for Computer Access Security System

A. Sermet Anagun

Eskişehir Osmangazi University, Industrial Engineering Department,
Bademlik 26030, Eskişehir, Turkey
sanagun@ogu.edu.tr

Abstract. A computer access security system, a reliable way of preventing unauthorized people for accessing, changing or deleting, and stealing the information, needed to be developed and implemented. In the present study, a neural network based system is proposed for computer access security for the issues of preventive security and detection of violation. Two types of data, time intervals between successive keystrokes during password entry through keyboard and voice patterns spoken via a microphone, are considered to deal with a situation of multiple users where each user has a certain password with different length. For each type of data, several multi-layered neural networks are designed and evaluated in terms of recognition accuracy. A committee neural network is formed consisting of six multi-layered neural networks. The committee decision was based on majority voting of the member networks. The committee neural network performance was better than the neural networks trained separately.

1 Introduction

A computer security system should not only be able to identify a person and let him/her access to the system if he/she has a correct security code or deny the access otherwise - *preventive security*, but also be capable of identifying the person whether he/she is indeed the right person - *detection of violations* [1]. To accomplish these goals, a software-based, a hardware-based, or a software and hardware-based security system may be used. In either case, although each person, who is eligible for accessing to the system, has his/her own security code, the code may be found with a trial-error process or stolen from the authorized person by someone else. When this occurs, the attempt made by a person may not be prevented. Due to the drawbacks of the common approaches, a different method in terms of computer access security, which may prevent copying or duplicating the security code issued, should be developed to differentiate an authorized person from the others such that valuable and/or more sensitive information for an organization should be secured.

Several researchers [2-7], in the area of computer access security, have concentrated on user identification based on individual's typing pattern, considered as a special characteristic for each person, using classical pattern recognition techniques, fuzzy algorithms, and neural networks (NNs) as powerful tools for pattern recognition and classification applications. In these studies, time intervals between successive keystrokes while entering a known and long password, an example of software-based

security system, on a keyboard were considered as an alternative security code to prevent unauthorized person for accessing the system involved and changing some information. Since the same password has been entered by a group of people, this situation may be classified as *multiple users-single password*. In addition, the studies mentioned have focused on preventive security, which basically classifies people into two groups; people who know the correct password and who do not know, without evaluating whether they are indeed authorized.

However, due to the developments in computer technology and the complexity of information systems, which the organizations might have, there may be different situations, which needed to be considered such as *multiple users-multiple passwords*, *single user-single password*, and *single user-multiple passwords*. As discussed in [8], each of these situations may be applied to the computer access security systems considering passwords with different lengths depending on his/her preferences or system's requirements, if applicable. They proposed a multi-layered NN based computer access security system for a situation where *multiple users-multiple passwords* with different lengths. In order to identify users and differentiate valid users from invalid ones (intruder), the NN was trained using a large set of data consisted of keystroke patterns of the participants. During the data collection process proposed, the users were asked to type their own passwords and the other passwords of the remaining participants. The designed system for the *multiple users-multiple passwords* case, has provided approximately 3% error and performed better than a statistical classifier based on Euclidean distance, 13.6%.

On the other hand, the computer access security system should be designed for not only the purpose of preventive security, but also the purpose of detection of violations to make the system more reliable. In the study of [9], a NN based system has been designed and applied to the cases of *multiple users-one password* and *multiple users-multiple passwords* with different lengths for preventive security and detection of violation purposes. Two critical issues, password-dependent identification (the lengths of the passwords different) and password-independent identification (the lengths of the passwords equal) were evaluated in terms of recognition accuracy. It has demonstrated that the users were classified or attempts of an intruder were denied 98.7% of the time.

A multi-layered NN for a computer access security system trained via voice patterns, spoken passwords through a microphone is designed by [1]. It has mentioned that based on their passwords, the users were recognized approximately at a value of 5.5% using the results of the designed of experiments for the NN's parameters and performed better than a statistical classifier. Recently, Anagun [10] proposed a two-stage procedure based on sequentially organized NNs for computer access security system.

Here, an intelligent computer access security system using a committee NN consisting of multi-layered NNs trained with a backpropagation learning algorithm is proposed for a situation of *multiple users-multiple passwords*. In order to differentiate authorized person from an intruder, the data composed of time intervals between keystrokes typed via keyboard and voice patterns spoken through a microphone are obtained by means of a data collection systems designed.

2 Data Collection

The time intervals between successive characters occurred, called keystroke dynamics, while entering a password using a keyboard, and finger prints and properties of voice of a person may be considered as person-dependent characteristics. These characteristics, also called special characteristics, may be somehow used in the form of a software-based system for user identification in or to differentiate users of a computer system to secure the information stored. In the present study, two different ways for differentiating a valid user from the others are used to find a better way for a computer access security system in terms of reliability. Based on the selected special characteristics, two types of data are collected from the same group: keystroke dynamics obtained during a password entry via a keyboard and voice patterns recorded as they are being spoken through a microphone.

In order to discuss whether a system mentioned may be applicable to computer access security, a network is formed composed of group of people and passwords with different lengths are assigned to the participants according to their preferences. Afterwards, each participant is asked to enter all of the passwords, using a keyboard and a microphone, respectively, in a random order and four times (arbitrarily between 9 A.M. and 5 P.M.) a day of each week for the period of three months. After the entries completed, both keystroke dynamics and the voice patterns for each user are evaluated and additional entries are made until the necessary number of patterns has been reached statistically. The data for the same passwords belong to the same persons are obtained in a different fashion.

2.1 Keystroke Dynamics

During the password entry process, the users, each of whom has different levels of computer skills, were asked to enter his/her own password and other passwords of the remaining members of the group, which are represented by “*” during the typing process, along with a user identification number in a random order based on a proposed data collection structure. After each entry, a typed phrase via keyboard is displayed on the bottom of the screen followed by a return key.

When the password is typed correctly, the time intervals between successive characters of the password being typed are computed and automatically recorded in a file according to the user and password identification numbers.

During this process, for instance, if the password of ENGINEERING is entered by the first member of the group, the time intervals of (E,N), (N,G), (G,I), (I,N), (N,E), (E,E), (E,R), (R,I), (I,N) and (N,G) would be computed and recorded in a file. Such a file, for each entry, consists of the time intervals of the password typed, user identification number that represents who typed the password, and password identification number that represents which password typed as follows:

$$T_1 T_2 T_3 \dots T_N \quad U_1 U_2 U_3 \dots U_K \quad P_1 P_2 P_3 \dots P_J$$

where,

T_i is the i^{th} time interval for the P_j^{th} password typed by the U_k^{th} user, $i = 1,2,3,\dots,N$

P_j is the j^{th} password typed by the U_k^{th} user (0 or 1), $j = 1,2,3, \dots,J$

U_k is the k^{th} user (0 or 1), $k = 1,2,3, \dots,K$

A recorded example pattern for the second password (ENGINEERING) entered via a keyboard by the first user is given as:

22 28 16 11 6 16 11 6 16 17 0 0 0 1 0 0...0 0 1 0 ...0

In order to process all the data obtained from the participants within the same NN structure, the time intervals of the shortest password are made equal to the dimension of the longest password by adding a necessary number of zeros.

2.2 Voice Patterns

Many different models have been postulated for quantitatively describing certain factors involved in the speech process. One of the most powerful models of speech behavior is the linear prediction model which has been successfully applied to the related problems in recent years [11].

In speech processing, a phrase is spoken into a microphone, recorded on audio tape as waveform, and then analyzed. The recorded speech waveform has a very complex structure and continually time-varying. The waveforms are analyzed based on frames (shifted windows along with the speech sequence). As the frames dynamically move through time, considering speech is dynamic and information-bearing process, transient features of the signal may be captured [12]. In order to capture the features of the signal, linear-invariant models over short intervals of time for describing important speech events should be implemented.

There are two well-known and widely used linear prediction models; the autocorrelation and the covariance methods. The autocorrelation method is always guaranteed to produce a stable linear prediction model [11]. The solution of the model is referred to as the autocorrelation method of determining the linear prediction coefficients (LPCs) or parameters [13]. The LPCs have been shown to retain a considerable degree of naturalness from the original speech. Thus, linear prediction models have been applied to speaker identification and verification.

During the password entry process, users are asked to speak the passwords clearly through a microphone and voice patterns of the passwords sampled at 8 bit and 16 kHz are recorded and digitized using WaveStudio. The recorded voice patterns are then transformed to LPCs using the autocorrelation method by means of Matlab to represent each voice pattern as frames or Hamming windows consisting of a certain number of data points and to reduce the dimension of each pattern. After transformation, the voice patterns are represented as follows:

$$X_1 X_2 X_3 \dots X_M \quad U_1 U_2 U_3 \dots U_K \quad P_1 P_2 P_3 \dots P_J$$

where,

- X_i is the i^{th} linear prediction coefficient of an Hamming window corresponding to the P_j^{th} password spoken by the U_k^{th} user, $i = 1,2,3,\dots,M$
- P_j is the j^{th} password spoken by the U_k^{th} user (0 or 1), $j = 1,2,3, \dots,J$
- U_k is the k^{th} user (0 or 1), $k = 1,2,3, \dots,K$

The recorded voice pattern for the second password (ENGINEERING) spoken through a microphone by the first user is depicted in Fig. 1.

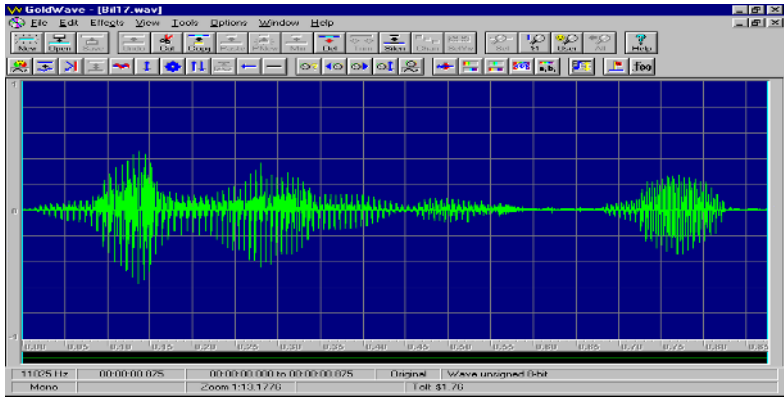


Fig. 1. Voice pattern for the password of ENGINEERING

3 Neural Network Architecture

It has been shown that a layered NN provides more potential alternatives than traditional pattern recognition techniques [1,8,14-15]. A pattern recognition technique, defined as a classification model, is concerned with performing feature extraction, learning the transparent mapping and classifying patterns [16]. For the task of pattern recognition using layered NNs, inputs correspond to features, connections between layers correspond to mapping, and outputs correspond to pattern classes. In addition, a layered NN may contain one or more hidden layers, which represent the domain knowledge and help to perform feature extraction. On the other hand, when a signal, a voice pattern, is represented in the frequency domain, as discussed in [17], signal-processing techniques can be used to determine the basic characteristics of the physical system involved. However, if a large number of examples can be obtained, NNs can be applied to eventually carry out the desired information or signal processing operation using these examples recorded either time or frequency domain [18]. Therefore, instead of signal processing techniques, the NN approach is mostly preferred for recognition and classification purposes.

In this study, two NNs, one for each type of data, are designed consisting of three layers; input, hidden, and output. Each layer is connected to the upper layer, inter-layer connections, via weights, randomly generated real values. A sigmoid function is used to determine the new activation values of the neurons in the hidden and output layers, respectively. Based on the results of the study of [19], the backpropagation algorithm is selected for training the designed NNs.

The number of neurons in the multi-layered NN architecture is varied depending on the data. The input layer is composed of 8-13 neurons, represented time intervals between successive keystrokes obtained from the passwords entered and 75-100, represented LPCs obtained from the transformation process for the passwords spoken. The number of neurons representing the user and typed/spoken password in the output layer are also varied depending on the experiments.

As indicated in [20], more neurons in the hidden layer reduce total error in training; however, fewer neurons increase the network performance in terms of

generalization, meaning that the ability to correlate a pattern with previously used patterns. For that reason, the number of neurons in the hidden layer, which yields to extract features between the input and the corresponding output pattern, are varied depending on the experiments to improve the network performance in regards to generalization. For the keystroke dynamics, the hidden neurons are varied in the range of (4-10), for the voice patterns, in the range of (30-50), respectively. The learning rate is assigned to 0.05, and momentum term to 0.3 based on designed experiments conducted by [1].

4 Experimental Results and Discussion

In regard to the computer access security system, several experiments are designed to investigate the overall performance for the system concerned. Each password is assigned to each user only based on his/her preferences. Then, the data consists of either time intervals or LPCs belong to a specific password selected by a user are introduced to a NN to initiate a *multiple users-multiple passwords* situation. This experiment is performed for both *preventive security* and *detection of violation* purposes. Since each user has a certain password to access to a part of or complete system, this situation may be considered as password-dependent recognition. The experiments are discussed in different sections and the results obtained are compared as follows.

4.1 Keystroke Dynamics Used

The collected data from password entry process were normalized time intervals based on the fraction of the largest element in the data set before presenting them to the NN. The data consisted of time intervals belong to a specific password selected by a user were introduced to a NN, which had N input and (K+J) output neurons.

The multi-layered NNs were trained using the proper data prepared for each of the experiments. In testing phase, the patterns which were not included in the training set, were fed to the designed NNs and the performance of the each NN was evaluated according to the correct/wrong classifications (Type I error). Time intervals obtained from an unauthorized person for the system involved and not included in training data were also tested to verify whether the person may be considered as an intruder (Type II error).

An overall recognition accuracy of 98.8% was obtained for training phase, and Type I and Type II errors were about 2.2% and 4.6%, respectively, in testing, since both user identification number and the pattern code of the entered password were examined simultaneously at each query. The results concluded that when the user identification number and a password for that user were questioned simultaneously, a better performance in computer access security system might be obtained.

4.2 LPCs Used

The same experiment was also conducted using LPCs in terms of preventive security and detection of violation. A three-layer NN was designed with the architectures of M input and K output neurons for the first experiment, M input and (K+J) output neurons for the second experiment, respectively. The designed NN was trained using the

normalized LPCs obtained by transforming the voice patterns. The training was maintained until a predetermined margin value was reached, then the performance of the NN was evaluated according to the Type I error. An overall recognition accuracy of 97.4% was obtained for training phase, and Type I and Type II errors were increased to 5.5% and 10.7%, respectively, in testing. Regarding with the results, the performance of this experiment was significantly lower than the previous one due to the drawbacks of the sampling procedure being used to record the voice patterns. In other words, if a security system were designed based on voice patterns, an intruder would be identified as valid user approximately 11% of the time. On the other hand, the system designed based on the keystroke patterns could not provide 100% accuracy as well.

According to the results, it has also observed that the users were easily and successfully identified and/or classified into proper groups when the sequence and placement of the characters appeared in the passwords are compatible in terms of vowel-consonant and distance between them, and the pronunciation of the passwords are appropriate in terms of linguistics.

4.3 Keystroke Dynamics and LPCs Used

Based on the results of the NNs consisted of different hidden neurons, the best three structures in terms of recognition accuracy are selected to configure a committee NN. A committee NN is then formed by recruiting six neural networks trained with different types of data into a decision-making process. The committee NN provides a reliable technique especially for speech based speaker verification when compared to a single network [21]. Addition, as mentioned in [22], a committee approach to classification is known to produce generally improved results, provided that error rates are less than 50% for each member of the committee.

In order to improve the reliability of the system concerned, a committee NN is design to be able to differentiate an authorized person from an intruder. The LPCs extracted from the speech signal and keystroke dynamics are fed to the committee NN. The decision is based on a simple majority opinion of the member networks. That is, the user may be allowed to access to the information stored if he/she is recognized by at least four out of six NNs (i.e. two of three NNs of each group producing the same results or making unanimous decision for the each attempt) as the same person. Otherwise, the attempt for the user would be rejected; thus, the information stored may be secured. The block diagram of the committee NN is shown in Fig. 2.

Both keystroke dynamics and LPCs type data obtained from the participants are used to examine the performance of the committee NN. Approximately 1.7% of the attempts are rejected by the committee NN (Type I error) due to the conflicting results obtained from the NNs trained with different types of data, although the person is authorized.

On the other hand, a tremendous reduction in the error value, approximately 4.2%, is obtained for the attempts of an intruder. That is, when special characteristics of the participants are evaluated based on a majority voting, the risk of accepting an attempt for a person although he/she is not authorized (Type II error) may be reduced. This concluded that, the committee NN had the ability to improve the reliability of the system as far as security is concerned.

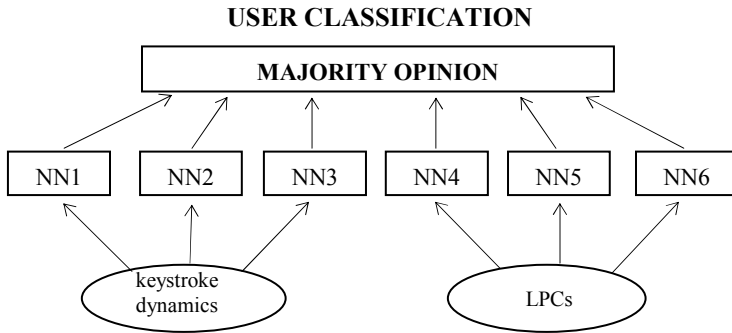


Fig. 2. Architecture of the committee NN. The first three networks are the best NNs trained with keystroke dynamics, whereas the last three networks are the best NNs trained with LPCs in terms of recognition accuracy.

5 Conclusion

In the present study, time intervals between successive keystrokes and voice patterns, which may be considered special characteristics of the users, are used to differentiate users, prevent unauthorized person to access the system, and try to detect the intruders by means of NNs. The experimental results showed that the NN trained using time intervals of a specific password along with the user identification number provided better performance.

It has observed that the data structure had a major effect on the performance of the network designed. The sequence and placement of the characters appeared in the passwords in terms of vowel-consonant and distance between them, and the pronunciation of the passwords in terms of linguistics are revealed to be considered for such a system.

Even if the voice patterns are considered as special characteristics for human beings, the NN trained by means of those patterns could not provide higher accuracy as expected. The voice pattern may be sampled at different parameters, although it increases the number of data points for each record and the training time of the neural network. The linear prediction model used in the study may be modified to produce Hamming windows consisting of more voice data to precisely capture the features of the signals. In order to improve overall performance of the system designed, a committee NN with majority voting was developed for computer access security system in the case of *multiple users-multiple passwords*. Based on the results, it has been observed that the committee NN was able to differentiate attempts made by the authorized person from an intruder with the accuracy of 98.3%, whereas 95.8% of the attempts made by an intruder were declined by the committee NN.

Other issues, such as seeking better security code alternatives (e.g. fingerprints, handwritten signatures, smart cards, and images) to differentiate users more precisely, investigating distinct NN architectures in terms of the number of neurons/layers, types of connections and the cases of *single user-single password*, *multiple users-single password*, and *single user-multiple passwords* to be able to implement this approach in an on-line mode are still available for further investigation.

References

1. Anagun, A.S.: An Artificial Neural Network Approach for a Computer Access Security System Based on the Characteristics of the Users. *Endüstri Mühendisliği*. 10 (1999) 3-11
2. Hussein, B.R., McLaren, R., Bleha, S.A.: An Application of Fuzzy Algorithms in a Computer Access Security System. *Pattern Recognition Letters*. 9 (1989) 39-43
3. Bleha, S.A., Slivinsky, C., Hussein, B.: Computer-Access Security Systems Using Keystroke Dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 12 (1990) 1217-1222
4. Bleha, S.A., Obaidat, M.S.: Dimensionality Reduction and Feature Extraction Application In Identifying Computer Users. *IEEE Transactions on Systems, Man, and Cybernetics*. 21 (1991) 452-456
5. Obaidat, M.S., Macchairolo, D.T., Bleha, S.A.: An Intelligent Neural Network System for Identifying Computer Users. *ASME Intelligent Engineering Systems through Artificial Neural Networks*. Ed. Dagli *et al.*, 1 (1991) 953-959
6. Bleha, S.A., Obaidat, M.S.: Computer Users Verification Using the Perceptron Algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*. 23 (1993) 900-902
7. Obaidat, M.S., Macchairolo, D.T.: An On-Line Neural Network System For Computer Access Security. *IEEE Transactions on Industrial Electronics*. 40 (1993) 235-242
8. Anagun, A.S., Cin, I.: An Alternative Way for Computer Access Security: Password Entry Patterns. *Proceedings of the 18th National Conference on Operations Research and Industrial Engineering*, Istanbul, Turkey. (1996) 17-20
9. Anagun, A.S., Cin, I.: A Neural Network Based Computer Access Security System for Multiple Users. *Computers and Industrial Engineering*. 35 (1998) 351-354
10. Anagun, A.S.: Designing a Neural Network Based Computer Access Security System: Keystroke Dynamics and/or Voice Patterns. *International Journal of Smart Engineering Design*. 4 (2002) 125-132
11. Markel, J.D., Gray Jr., A.H.: *Linear Prediction of Speech*. Springer-Verlag, New York (1982)
12. Deller Jr., J.R., Proakis, J.G., Hansen, J.H.L.: *Discrete-Time Processing of Speech Signals*. Macmillian Publishing Co., New York (1993)
13. Rabiner, L.R., Schafer, R.W.: *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs (1978)
14. Burr, D.J.: Experiments on Neural Net Recognition of Spoken and Written Text. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 36 (1988) 1162-1168
15. Huang, W., Lippmann, R.: Comparisons between Neural Networks and Conventional Classifiers. *Proceedings of the 1st International Conference on Neural Networks*, (1987) 485-494
16. Pao, Y.H.: *Adaptive Pattern Recognition and Neural Networks*. Addison-Wesley, Reading (1989)
17. Freeman, J.A., Skapura, D.M.: *Neural Networks: Algorithms, Applications, and Programming Techniques*. Addison-Wesley Publishing Co., Reading (1991)
18. Soucek, B.: *Neural and Concurrent Real-Time Systems - The Sixth Generation*. John Wiley-Sons., New York (1989)
19. Anagun, A.S.: A Multilayered Neural Network Based Computer Access Security System: Effects of Training Algorithms. *Lecture Series on Computer and Computational Sciences*. 4B (2005) 1604-1607

20. Klimasauskas, C.C.: Applying Neural Networks, Part III: Training a Neural Network. *PC AI*. (1991) 20-24
21. Reddy, N.P., Buch, O.A.: Speaker Verification Using Committee Neural Networks. *Computer Methods and Programs in Biomedicine*. 72 (2003) 109-115
22. Jerebko, A.K., Malley, J.D., Franaszek, M., Summers, R.M.: Multiple Neural Network Classification Scheme for Detection of Colonic Polyps in CT Colonography Data Set. *Academic Radiology*. 10 (2003) 254-160

C-TOBI-Based Pitch Accent Prediction Using Maximum-Entropy Model

Byeongchang Kim¹ and Gary Geunbae Lee²

¹ School of Computer & Information Communications Engineering,
Catholic University of Daegu, South Korea

bckim@cu.ac.kr

² Department of Computer Science & Engineering,
Pohang University of Science & Technology, Pohang, South Korea

gblee@postech.ac.kr

Abstract. We model Chinese pitch accent prediction as a classification problem with six C-ToBI pitch accent types, and apply conditional Maximum Entropy (ME) classification to this problem. We acquire multiple levels of linguistic knowledge from natural language processing to make well-integrated features for ME framework. Five kinds of features were used to represent various linguistic constraints including phonetic features, POS tag features, phrase break features, position features, and length features.

1 Introduction

Assigning the appropriate pitch accent in text-to-speech systems is important for naturalness and intelligibility. As widely known, Chinese is a tonal language and a syllable is normally assigned as the basic prosody element in processing. Each syllable has a tone and a relatively steady pitch contour. However, the pitch contour is transformed from the isolated ones to influence each other when they appear in the spontaneous speech in accordance with different contextual information. Therefore, Chinese pitch accent prediction is considered as a complicated problem.

Many machine learning techniques have been introduced to predict pitch accent, including the Hidden Markov Model, neural network, decision trees, bagging, and boosting. Xuejing Sun proposed an ensemble decision tree approach in four-class English pitch accent label [8]. Their method shows high accuracy of 80.50% when they just used text features. The performance improvement is 12.28% for the baseline accuracy. Michell L. Gregory and Yasemin Altun proposed a CRF based approach in a two-class English pitch accent label [4]. They report an accuracy of 76.36%. The performance improvement is 16.86% for the baseline accuracy.

C-ToBI as an intermediate representation, normally increases system-level modularity, flexibility and domain/task portability, but should be implemented with no performance degradation. However, labeling C-ToBI on speech corpus is normally very laborious and time-consuming, so we used an automatic C-ToBI labeling method.

We treat entire pitch accent prediction as a classification problem and apply a conditional maximum entropy (ME) model based on C-ToBI tone and intonation label, which consists of 8 different classes including boundary tones. Various kinds of linguistic information are represented in the form of feature, and the five kinds of features we used in our system include phonetic features, POS tag features, phrase break features, position features, and length features.

2 Conditional Maximum Entropy (ME) Framework

In ME modeling, relations between input information and output information are expressed as feature functions. The model whose entropy is maximized under the constraints defined by feature functions must be estimated. It means that the model would be close to the uniform distribution for input information, which rarely appears in the training corpus. Chinese pitch accent prediction must relate many kinds of linguistic features from Chinese pitch accent annotated corpus and ME modeling can be effective for various feature integration.

A classifier obtained by means of an ME technique consists of a set of parameters that are estimated using an optimization procedure. Each parameter is associated with one feature observed in the training data. The main purpose is to obtain the probability distribution that maximizes the entropy, that is, maximum ignorance is assumed and nothing apart from the training data is considered. In addition to its effectiveness dealing with the sparse data problem, another advantage using the ME framework is that even knowledge-poor features can be estimated accurately by asking only elementary questions to the surrounding contexts. So, we use ME modeling for our Chinese pitch accent prediction task.

The ME model allows experimenters to encode various dependencies freely in the form of features [1]. Let us assume a set of contexts as X and a set of classes as C . The system chooses the class C with the highest conditional probability in the context X . The conditional probability can be calculated as equation 1, where K is the number of features and $Z(X)$ is a normalization factor that takes the form of equation 2. Each parameter λ_j corresponds to one feature f_j and can be interpreted as a weight for that feature. Therefore, we can say $p(c|x)$ is a normalized product of those features with their weights.

$$P(c|x) = \frac{1}{Z(x)} \exp\left(\sum_{j=1}^k \lambda_j f_j(c, x)\right) \quad (1)$$

$$Z(x) = \sum_c \exp\left(\sum_{j=1}^k \lambda_j f_j(c, x)\right) \quad (2)$$

To estimate the parameters λ_j , the maximum likelihood (ML) estimate can be shown to satisfy the constraints:

$$\sum_x \tilde{P}(x) \cdot \sum_c P(c|x) \cdot f_j(x, c) = E_{\tilde{P}} f_j(x, c) \quad (3)$$

where \tilde{P} is the empirical distribution of the given training corpus. If the constraints f_j is consistent, there exists a unique solution called a minimum discrimination information (MDI) solution or ME solution when the prior is uniform. This unique ME/MDI solution can be found by several iterative methods such as generalized iterative scaling (GIS), improved iterative scaling (IIS) or the limited-memory variable metric method, which is a limited-memory version of the quasi-newton method (also called L-BFGS). The L-BFGS was the most effective training method for the ME model [7]. We used the L-BFGS method to estimate the parameters of an ME model for our pitch accent prediction. The ME can be viewed as an ML training for exponential models, and, like other ML methods, is prone to overfitting of training data. We adopt a Gaussian prior smoothing from several proposed methods for our ME models. The Gaussian prior is a powerful tool for smoothing general ME models, and can work well in the language models [3]. This has the effect of changing equation 3 to

$$\sum_x \tilde{P}(x) \cdot \sum_c P(c|x) \cdot f_j(x, c) = E_{\tilde{P}} f_j(x, c) - \frac{\lambda_j}{\sigma_j^2} \quad (4)$$

for some suitable variance parameter σ_j^2 .

3 Chinese Pitch Accent Prediction

3.1 Chinese Intonation and C-ToBI

In the early 20th century, tone and intonation research for Chinese entered into a new phase due to two phoneticians: Dr. Liu Fu (Ban-nong) and Dr. Chao Yuanren (Y.R.Chao). Chao pointed out that syllabic tone patterns could be modified by the sentential attitudinal intonation, just like “the small ripples riding on top of large waves”. It clearly explains the relation between syllabic tone patterns and the sentential intonation contours.

Lexical tones in Chinese are known for their sliding pitch contours. When produced in isolation, these contours seem well defined and quite stable. When produced in context, however, the tonal contours undergo certain variations depending on the preceding and the following tones. We used C-ToBI labeling system to represent Chinese tonal features. C-ToBI is a TOBI-like transcription system developed for Chinese prosodic annotation. Table 1 shows the Tone and intonation labels in C-ToBI¹.

3.2 Pitch Accent Prediction Architecture

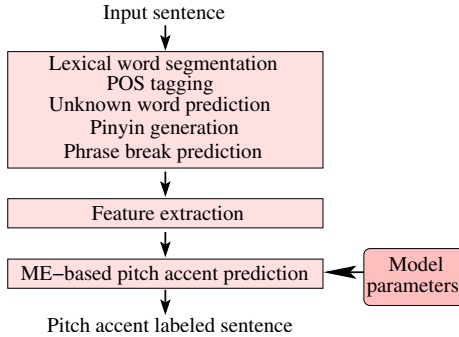
Fig. 1 shows a pitch accent annotation example. The overall architecture of our ME-based pitch accent prediction process is shown in the Fig. 2. In our research, we used previously developed word segmentation and POS (part-of-speech) tagging system called POSTAG/C [5], an unknown word prediction

¹ http://www.cass.net.cn/chinese/s18_yys/yuyin/english/ctobi/ctobi.htm

Table 1. Tone and intonation labels in C-ToBI (version 2.0)

Labels	Meaning
H-H, L-H, H-L, L-L	Tonal features for tones
H, L	Neutral tone
L%, H%	Boundary tones
\wedge , $\wedge\wedge$, $!$, $!!$	upstep, wide upstep, downstep, wide downstep
R	Register shifting

许多(m)/ B_0 电影(n)/ B_0 人(n)/ B_1 对此(d)/ B_1 也(d)/ B_1 都(d)/ B_0
 有(v)/ B_0 一些(m)/ B_1 议论(v)/ B_3 .
 xu:3(L-H) duo1(H-H) dian4(H-H) ying3(L-L) ren2(L-L)
 du14(L-L) ci3(L-L) ye3(L-L) dou1(H-H) you3(H-L)
 yi4(L-) xie1(L-L) yi4(L-L) lun4(L-L).

Fig. 1. Pitch accent labeled example**Fig. 2.** Pitch accent prediction process

system [6], a Chinese pinyin generation system [9] and a phrase break prediction system [11]. Using the previous linguistic analysis systems, we can extract five kinds of linguistic features, such as phonetic features, POS tag features, phrase break features, position features, and length features. Then we use the conditional maximum entropy model parameters estimated by L-BFGS [10] method to predict the pitch accent.

3.3 Multiple Linguistic Features

The features used in our model can be divided into two groups: isolated feature group and co-occurrence feature group.

Isolated feature group: The following uni-gram features are used.

- Phonetic features: Although the tonal contours undergo certain variations depending on the preceding and the following tones, the phonetic features still contain great prosody information and phonetic information is one of the most widely use predictor for pitch accent. Phonetic features include pinyin of current syllable, left syllable’s pinyin, right syllable’s pinyin, current syllable’s consonant, current syllable’s vowel, and current syllable’s tone. There include 38 different consonants, 21 different vowels, 5 different tones, and 1,231 different PinYins in our synthesis DB.
- POS tag features: POS tag information is one of the most widely used predictors in prosody modeling. A word might differ from itself in part-of-speech (POS), which carries various prosodic information. In this paper, POS is divided into 43 categories to capture the variations of prosodic features. POS tag features include current, previous and next POS tags.
- Phrase break features: Phrase break features include non-break, prosodic word, prosodic phrase, and intonation phrase.
- Position features: Position features include the position of a syllable in a sentence, a phrase and a word respectively. In general, the pitch contours in sentence, phrase and word will follow an intonation pattern. For example, the F0 contour will decline in a declarative sentence. This implies that the syllable position in a sentence, phrase and word will affect the prosodic information.
- Length features: Length features include current word length, next word length and sentence length.

Co-occurrence feature group: The following co-occurrence feature pairs are used.

- Current POS tag - position of syllable in a word.
- Current phrase break - position of syllable in a word
- Current syllable’s pinyin - position of syllable in a word
- Current POS tag - next POS tag
- Left syllable’s pinyin - current syllable’s pinyin
- Current syllable’s pinyin - right syllable’s pinyin
- Left syllable’s pinyin - current syllable’s pinyin - right syllable’s pinyin (triple feature)

3.4 Fundamental Frequency Contour Generation

In most synthesizers, the task of generating a prosodic tone using the ToBI label system consists of two sub-tasks: the prediction of intonation labels from input text, and the generation of a fundamental frequency contour from those labels and other information. We used a popular linear regression method to generate fundamental frequency from C-ToBI labels [2]. This method does not require any other rules for label types, and is general enough for many other languages. Our prediction formula is as follows:

$$target = w_1 f_1 + w_2 f_2 + \dots + w_n f_n + I \quad (5)$$

Where each f_i is the feature that contributes to the fundamental frequency, and we can decide the weights $w_1 \sim w_n$ and I through simple linear regression. We applied the above formula to every syllable and obtained a target value of the fundamental frequency. We prepared pitch values extracted from a speech file, divided them into five sections for each syllable, and predicted the fundamental frequency at every point.

4 Experimental Results

4.1 Corpus

The experiments are performed on a commercial Chinese database provided by Voiceware Inc. The database has 2,197 sentences, 52,546 Chinese characters, which consist of 25,974 Chinese lexical words. The database is POS tagged, pinyin annotated, break-labeled with four class prosodic structures, and pitch accent labeled with six classes. The occurrence probabilities of tones are shown in Table 2. We divide the database into 10 parts and conducted a 10-fold cross validation.

Table 2. Occurrence probabilities of tones in the corpus

Pitch Accent Classes	H-H	H-L	L-L	L-H	H	L
Occurrence Probabilities	16.0%	11.7%	43.4%	22.8%	2.2%	3.9%

4.2 Pitch Accent Prediction Experiment

We performed three experiments to show the pitch accent prediction results of our ME-based method. In the first experiment, we used isolated feature combinations to show the best feature selection. In the second experiment, we used co-occurrence feature combinations to show the best combination feature selection. In the third experiment, we set several Gaussian priorities for smoothing the maximum entropy models.

Table 3. Performance in each isolated features

Used Features	Acc (%)
Phonetic: PinYin(-1, 0, 1)	64.67
POS tag(-1, 0, 1)	45.63
Phrase break	43.37
Numeric	44.15

Table 3 shows the performance of each isolated feature class. The performance measure was simply defined as:

$$Acc = \frac{c}{N} * 100\% \quad (6)$$

Table 4 shows the accumulated performance of the best feature selection in each isolated feature class. As the experiment shows, we can draw various conclusions on the effect of isolated feature selection for C-ToBI-based pitch accent prediction.

Table 4. Accumulated performance in the best feature selection

	Used Features	Acc (%)
Base	Based on tone	35.51
Line	Based on each class frequency	43.40
Phonetic	Current pinyin	60.36
	Previous and next pinyin	64.67
Features	Consonant, vowel, and tone	65.43
POS	Current POS tag	65.75
Tag	Previous and next POS tag	66.96
	Phrase break	67.42
	Position of syllable in a sentence	67.68
Length	Current and next word length	67.70
	Sentence length	67.73

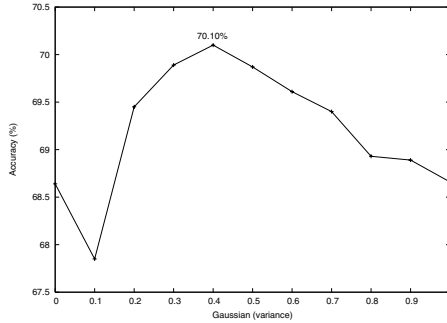
- The phonetic features are the most important features, which include current pinyin, previous pinyin, next pinyin, consonant, vowel, and tones.
- POS tag is a useful feature and the window size of 3 is the best in our experiment.
- Adding a phrase break feature is helpful.
- Position features and length features are only slightly useful for the pitch accent prediction.

Table 5 shows the result of co-occurrence feature selection. In this experiment, we can show that the co-occurrence features are also useful for pitch accent prediction. However, adding the last two co-occurrence features actually decreases the performance, because of the data sparseness. Nevertheless, through the Gaussian smoothing, we can get the best performance by finally adding Left pinyin - current pinyin - right pinyin co-occurrence features. Fig. 3 shows the Gaussian smoothing results. A Gaussian priority of 0.4 gives the best performance of 70.10% in our experiment.

As shown in Table 6, the improvement of performance is impressive compared with the previous systems, especially when we generate all the training corpora by automatic tone-labeling process. The improvements are significantly useful for applying to Chinese TTS systems.

Table 5. Performance in co-occurrence feature selection

Used Features	Acc (%)
Current POS tag - position in a word	68.60
Current phrase break - position in a word	68.68
Current syllable's pinyin - position in a word	68.78
Current POS tag - next POS tag	68.80
Left pinyin - current pinyin	68.88
Current pinyin - right pinyin	68.81
Left pinyin - current pinyin - right pinyin	68.64
Left pinyin - current pinyin - right pinyin (after Gaussian smoothing)	70.10

**Fig. 3.** Gaussian smoothing**Table 6.** Performance improvement

Class Number	Baseline (%)	Acc (%)	Improvement (%)
6	43.40	70.10	26.70

4.3 Fundamental Frequency Generation Experiment

We predicted five pitch values for each syllable and generated fundamental frequency with the interpolated method based on the predicted values. We detected pitch values per 16ms from speech with esps/Xwaves. For training and testing, we extracted a feature set from an automatically created file that is aligned with pitch sequences and phone sequences. Further, we eliminated items with '0' from pitch values as a noise for the constructing model.

We used seven category features to construct a linear regression model as the following features with pitch accent that we have predicted using C-ToBI representation.

- Current syllable’s consonant;
- Current syllable’s vowel;
- Current syllable’s tone;
- Current syllable’s POS tag;
- Current syllable’s phrase break index;
- Syllable’s position in current sentence;
- Pitch accent label.

Since we extracted five pitches from each syllable, we constructed a linear regression model for each pitch, and obtained a total of five models. Since the predicted response is a vector, we computed the RMSE² and correlation coefficients³ for each element of the vectors and averaged them. Table 7 shows the result of our generation model with C-ToBI and its comparison with direct generation without C-ToBI representation. Table 7 shows that we can improve the fundamental frequency generation performance using the C-ToBI based prediction model.

Table 7. The result of fundamental frequency generation

	Without C-ToBI	Using C-ToBI	Improvement
RMSE	47.920	40.552	-7.368
Correlation Coefficient	0.531	0.621	+0.090

5 Conclusions

We proposed a conditional maximum entropy model for Chinese pitch accent prediction task. We analyzed several combinations of linguistic features in order to identify which features are the best candidates for ME-based pitch accent prediction. Moreover, we generate the training corpora using fully automatic tone-labeling process. The results obtained from our proposed system show that

² The RMSE is a simple distance measure between two F0 contours that can be calculated as equation:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\tilde{P}_i - P_i)^2} \quad (7)$$

³ The correlation coefficient indicates whether the two contours exhibit the same tendencies by measuring the deviation from the mean F0 at regular set points in time that can be calculated as equation:

$$Corr = \sum_{i=1}^n \frac{(\tilde{P}_i - \tilde{P}_{mean})(P_i - P_{mean})}{\sqrt{(\tilde{P}_i - \tilde{P}_{mean})^2 * (P_i - P_{mean})^2}} \quad (8)$$

the selected best feature sets guarantee the success of the prediction method. Because the ME model allows experimenters to encode various dependencies freely in the form of features and often Chinese pitch accent related features are dependent on each other, the ME model's feature selection is more flexible than that of other machine learning models. As shown in our results, the performance of prosody generation using the C-ToBI system is better than the performance without the C-ToBI system.

References

1. A.L.Berger, S.A.D.Pietra, and V.J.D.Pietra, "A maximum entropy approach to natural language processing", Computational Linguistics, vol.22, no.1, 1996.
2. A.W.Black and A.J.Hunt, "Generating F0 contours from ToBI labels using linear regression", In proceeding of the international conference on spoken language processing(ICSPL), CSLI, 1996.
3. Stanley F. Chen and Ronald Rosenfeld, "A Gaussian Prior for Smoothing Maximum Entropy Models", Technical Report CMU-CS-99-108, 1999.
4. Michelle L. Gregory, Yasemin Altun, "Using conditional random fields to predict pitch accents in conversational speech", ACL, 2004.
5. Ju-Hong Ha, Yu Zheng, Gary G. Lee, "Chinese segmentation and POS-tagging by automatic POS dictionary training", In Proceedings of the 14th Conference of Korean and Korean Information Processing, 2002.
6. Ju-Hong Ha, Yu Zheng, Gary G. Lee, "High speed unknown word prediction using support vector machine for Chinese Text-to-Speech systems", IJCNLP, 2004
7. Robert Malouf, "A comparison of algorithms for maximum entropy parameter estimation", In proceedings of CoNLL-2002, 49-55, Taipei, Taiwan, 2002.
8. Xuejing Sun, "Pitch accent prediction using ensemble machine learning", ICSLP, 2002.
9. Hong Zhang, JiangSheng Yu, WeiDong Zhan, and ShiWen Yu, "Disambiguation of Chinese polyphonic characters", International Workshop on Multimedia Annotation, 2001.
10. Zhang Le, "Maximum entropy modeling toolkit for python and C++", <http://www.nlplab.cn/zhangle/>, 2003.
11. Yu Zheng, Byeongchang Kim, Gary Geunbae Lee, "Using multiple linguistic features for Mandarin phrase break prediction in maximum-entropy classification framework", ICSLP, 2004.

Design and Fabrication of Security and Home Automation System

Eung Soo Kim¹ and Min Sung Kim²

¹ Div. of Digital Information Engineering,
Pusan University of Foreign Studies,
55-1, Uam-dong, Nam-Gu, Busan, 608-738, Korea
eskim@pufs.ac.kr

² Dept. of Information & Communications Engineering,
TongMyong University of Information Technology,
535 Yongdang-dong, Nam-Gu, Busan, 608-711, Korea
minsung@tit.ac.kr

Abstract. Home automation system was designed and fabricated for controlling of home appliances, gas detection and home security. The fabricated system could detect the intruder using infrared sensor and monitor the room or office in real time by web camera. The password was needed to enter the house or office and the operation of home appliances could be remotely controlled by network, too. This fabricated system was small and had an advantage to supplement additional function easily.

1 Introduction

Nowadays the increasing developments of home automation and security system have been driven primarily by the need to promote the benefit of our lives. Home automation is needed to maintain safety, convenience, and comfortableness at home as controlling the home appliances, detection of gas leakage, fire alarm, controlling the light, and monitoring the visitors. Some conditions, however, are necessary for security and home automation system. First, the power consumption of home automation system should be low and the size must be small because it could be used in house. Second, the installation and maintenance of the system should be easy. Finally, it is easy to operate the system and could be controlled remotely through the Internet. The concern about security system has been increasing to protect the human beings and valuable things. A sensor is important in security system. Infrared sensor, thermo sensor, optical sensor, and biometrics such as finger print recognition and iris recognition are usually used in security system[1-6]. In this paper, security and home automation system are designed by VHDL and fabricated using CPLD[7]. The fabricated system was small and showed good performance. In addition to, we could add other functions easily.

2 System Design and Fabrication

The flowchart of security and home automation system is shown in Fig. 1. Architecture was made by VHDL and we verified the model with VHDL simulator.

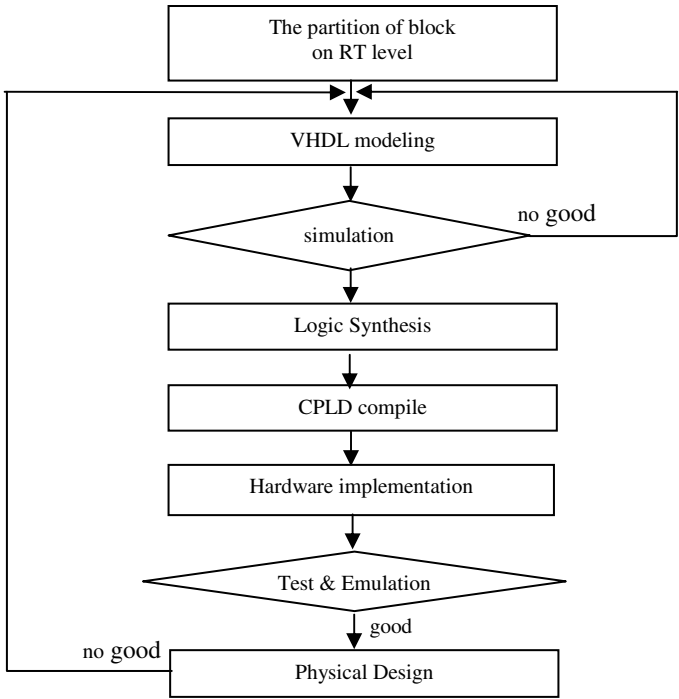


Fig. 1. Flowchart of the security and home automation

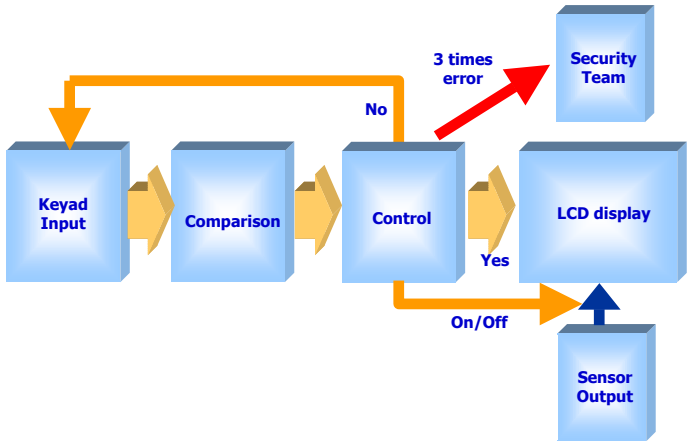


Fig. 2. Block diagram of door control in security system

For security system, we used two infrared sensors to detect an intruder and the keypad to permit a visitor to enter a house or office. The block diagram of door control in security system is shown in Fig. 2. Keypad input is used as an input signal. If a

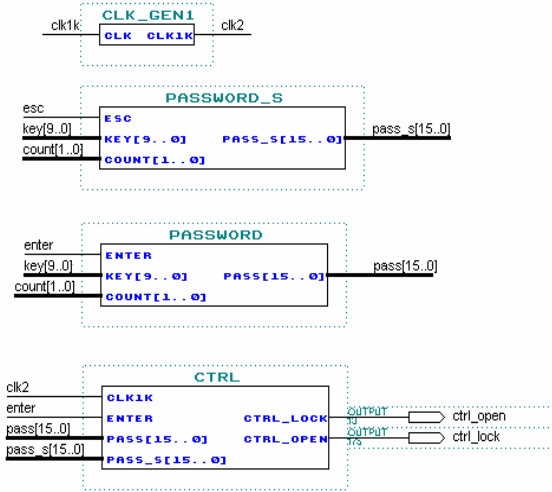


Fig. 3. The part of keypad input

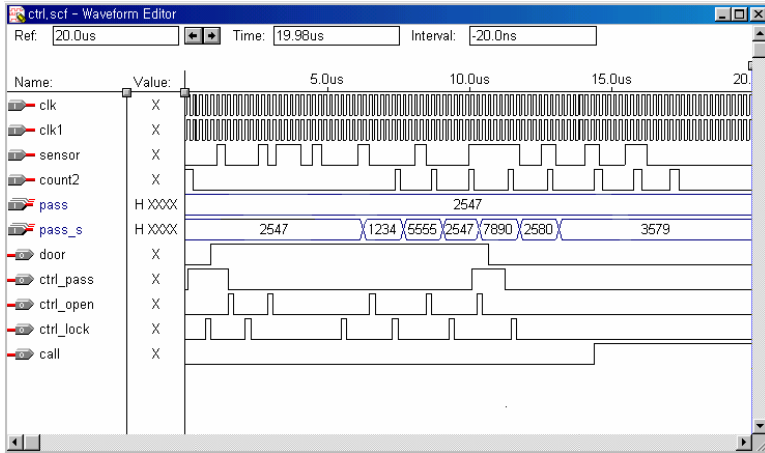


Fig. 4. The simulation of sensor and door control part

Table 1. The distance between photo detectors

	The distance between the infra-red sensor and the photo detector	The distance between photo detectors
The distance between top and bottom	13.5 cm	5.48 cm
The distance between left and right	7.5 cm	3.6 cm

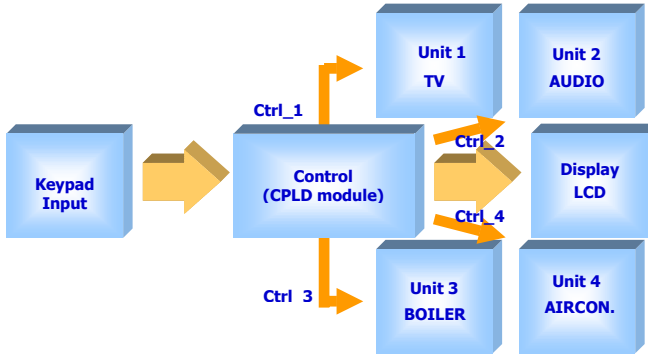


Fig. 5. The control part of the home automation

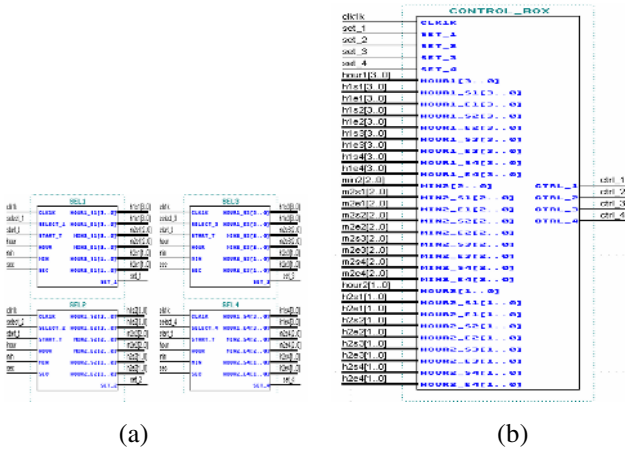
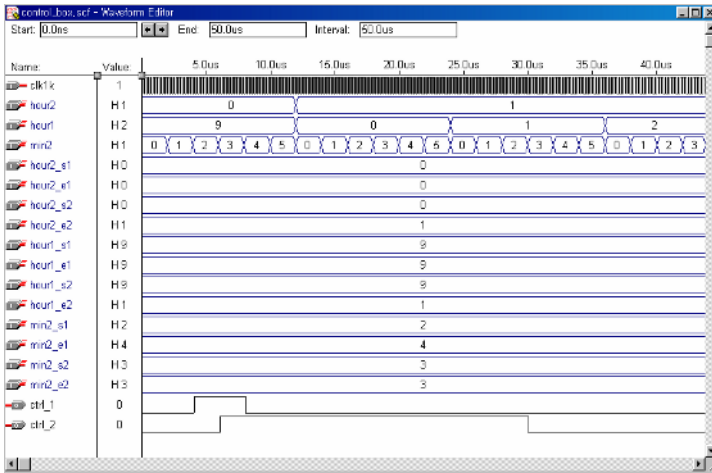


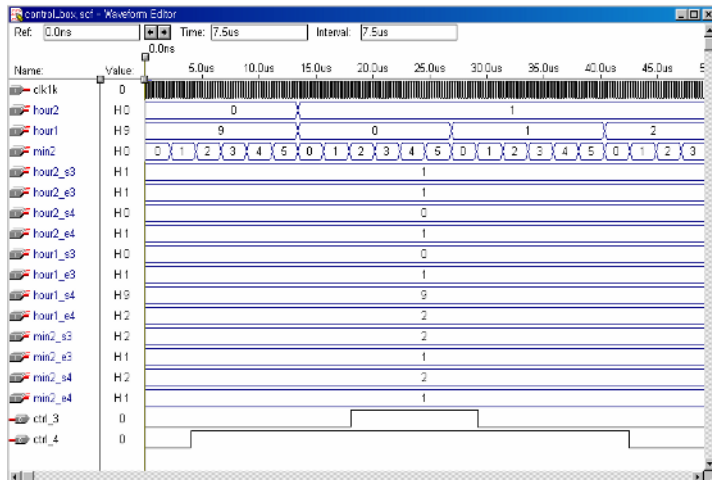
Fig. 6. The input and output signals of CONTROL_BOX

keypad input were the same as password saved in security system, the door would be opened. Otherwise, the door would not be opened. If input password is wrong in 3 times, the system informs policeman or host of alarm signal. Fig. 3 shows the part of keypad input. The CTRL compares the password with keypad input and synchronized the clock frequency. We could change the password anytime in 5 seconds. Two infrared sensors were used to detect an intruder at the front of the door. The sensor signals were compared and controlled the door. The simulation of sensor and door control part is shown in Fig. 4, where pass and pass_s are the password saved by host and password inputted by visitor, respectively. The distance between two photo detectors receiving the infrared signal is investigated to improve the sensitivity of photo detectors when the distance between the infrared source and the photo diode is constant because the radiation angle of the infrared source is 27° at operation current 50mA. The results are shown in Table 1. Home automation system controls home appliances such as TV, audio component, air conditioner, lights and web camera in computer and

detects gas leakage. In addition, remote control of the home appliances by network is possible so that a host can reserve the start time and stop time of them. The host can monitor the room or office in the web camera in anytime and anyplace through the Internet and move the web camera all directions. This fabricated home automation system can easily extend additional home appliances, too. The block diagram of the control part of home automation is shown in Fig. 5. We could choose the each home appliance to set the operation time and LCD panel displays the states of it. The control signals of each appliances are shown in Fig. 6(a), where SEL1, SEL2, SEL3, and SEL4 represent home appliances, respectively. The operating time of each component was inputted in CONTROL_BOX and they were operated according to setting time



(a)



(b)

Fig. 7. The simulations of CONTROL_BOX

which is output signal from the CONTROL_BOX as shown in Fig. 6(b). Fig. 7 shows the simulation of CONTROL_BOX. The control signals (ctrl_1, ctrl_2, ctrl_3 and ctrl_4) of SEL1, SEL2, SEL3, and SEL4 were generated at set time. In Fig. 7 the hour2, hour1 and min2 represent present time, s1 (s2, s3, and s4) and e1 (e2, e3, and e4) are start time and stop time of SEL1 (SEL2, SEL3, and SEL4). All home appliances are well operated by the set time.

3 Operation of the Systems

Security and home automation system are fabricated using CPLD (Altera Inc.) and tested. 56 % of total cell amount was used. Fig. 8 is LCD panel and keypad, where LCD panel indicates the door states, input password signal with asterisk marks, and phone number of police office or host which the system call to unless the visitor enters correct password in 3 times. This LCD panel also displays the operating time of home appliances. Fig. 9 shows the monitor of remote computer, which displays the states and operating time of home appliances and room or office image by the web camera. We can choose the home appliances, set the operating time of the home appliances, and set the temperature of air conditioner and heater through the network. And we can move the web camera to see the house or office anytime and anyplace by the network as we click the control button displayed in computer monitor. The system showed good operation. When the control program written in C++ was terminated, the computer saved the latest data automatically.

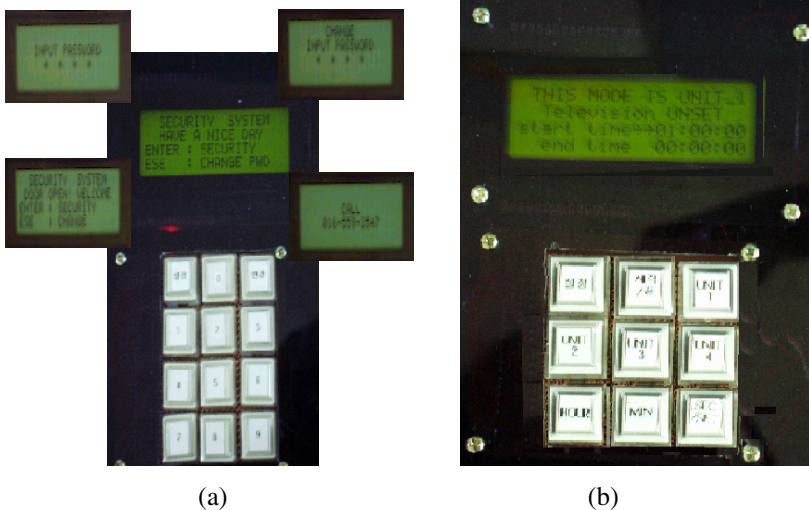


Fig. 8. LCD panel and keypad of the fabricated system. (a) security states and (b) the state of TV of home automation system.

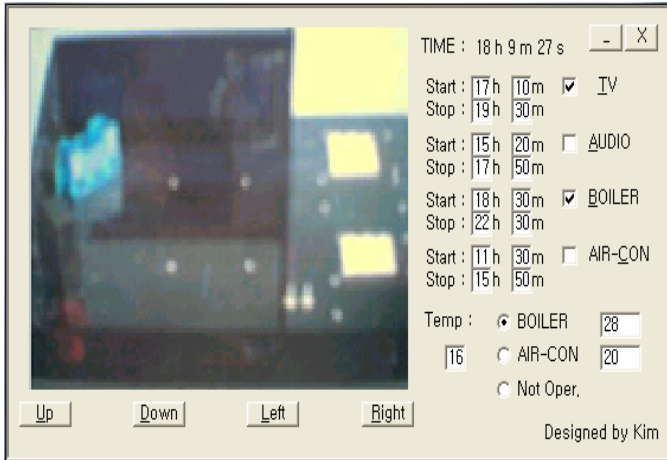


Fig. 9. The room monitored by web camera and the states of home appliances in computer monitor

4 Conclusions

The needs of Home automation and security system have been increasing for comfortableness and protecting human and properties. We have designed the security and home automation system by VHDL and fabricated the system using CPLD. The fabricated system informs you of emergency alarm when it detects intruders. Home automation system remotely controls electronic through network. This system was small and can be extended easily both additional functions and other home appliances.

References

1. Korsah, K., Kisner, R., Boatner, L., Christne, H., Paris, D.: Sensors and Actuators. A : Phys. **1119** (2005) 358-364
2. Zhang, X., Dalsgaard, E., Liu, S., Lai, H., Chen, J.: Appl. Opt. **36** (1997) 8096-8097
3. Porwik, P., Wieclaw, L.: IEICE Elec. Exp. **1** (2004) 575-581
4. Mardia, K. V., Boczokowsky, A. J., Feng, X., Hinsworth, T. J.: Pattern Recognition Lett. **18** (1997) 1197-1203
5. Boles, W. W.: Eng. Appl. Of Artif. Int. **11** (1998) 77-85
6. Narayanswamy, R., Johnson, G. E., Silveira, P. E. X., Wach, H. B.: Appl. Opt. **44** (2005) 701-712
7. Roth, C. H. Jr.: Digital system design using VHDL, Thomson Learning (1998)

PGNIDS(Pattern-Graph Based Network Intrusion Detection System) Design*

Byung-kwan Lee, Seung-hae Yang, Dong-Hyuck Kwon, and Dai-Youn Kim

Dept of Computer Engineering, Kwandong University, Korea
bklee@kd.ac.kr, yang7177@cho.com, taz0108@hotmail.com,
dy2300@freechal.com

Abstract. PGNIDS(Pattern-Graph based Network Intrusion Detection System) generates the audit data that can estimate intrusion with the packets collected from network. An existing IDS(Intrusion Detection System), when it estimates an intrusion by reading all the incoming packets in network, takes more time than the proposed PGNIDS does. As this proposed PGNIDS not only classifies the audit data into alert and log through ADGM(Audit Data Generation Module) and stores them in the database, but also estimates the intrusion by using pattern graph that classifies IDPM(Intrusion Detection Pattern Module) and event type, Therefore, it takes less time to collect packets and analyze them than the existing IDS, and reacts about abnormal intrusion real time. In addition, it is possible for this to detect the devious intrusion detection by generating pattern graph.

1 Introduction

PGNIDS is proposed to design network intrusion detection system based on pattern graph. When information is exchanged, PGNIDS can detect illegal connectivity and intrusion-related behavior such as DoS(Denial of Service) attack and port scan.

2 Related Works[12]

2.1 IDS Analysis

There are two primary approaches to analyzing events to detect attacks: misuse detection and anomaly detection. Misuse detection, in which the analysis targets something known to be "bad", is the technique used by most commercial systems. Anomaly detection, in which the analysis looks for abnormal patterns of activity, has been, and continues to be, the subject of a great deal of research. Anomaly detection is used in limited form by a number of IDSs.

* This work was supported by grant No. B1220-0501-0315 from the University fundamental Research Program of the Ministry of Information & Communication in Republic of Korea.

2.2 The Kind of Intrusion Detection

Some IDSs analyze network packets captured from network backbones or LAN segments, to find attackers. Other IDSs analyze sources information generated by the operating system of application software for signs of intrusion.

2.2.1 Network-Based IDSs

The majority of commercial intrusion detection systems are network-based. These IDSs detect attacks by capturing and analyzing network packets. Listening on a network segment or switch, the network-based IDS can monitor the network traffic affecting multiple hosts that are connected to the network segment, thereby protecting those hosts. Network-based IDSs often consist of a set of single-purpose sensors or hosts placed at various points in a network. These units monitor network traffic, performing local analysis of that traffic and reporting attacks to a central management console. As the sensors are limited to running the IDS, they can be more easily secured against attack. Many of these sensors are designed to run in "stealth" mode, in order to make it more difficult for an attacker to determine their presence and location.

2.2.2 Host-Based IDSs

Host-based IDSs operate on information collected from within an individual computer system. This vantage point allows host-based IDSs to analyze activities with great reliability and precision, determining exactly which processes and users are involved in a particular attack on the operating system. Furthermore, unlike network-based IDSs, host-based IDSs can "see" the outcome of an attempted attack, as they can directly access and monitor the data files and system processes usually targeted by attacks.

2.2.3 Application-Based IDSs

Application-based IDSs are a special subset of host-based IDSs that analyze the events transpiring within a software application. The most common information sources used by application-based IDSs are the application's transaction log files. The ability to interface with the application directly, with significant domain or application-specific knowledge included in the analysis engine, allows application-based IDSs to detect suspicious behavior due to authorized users exceeding their authorization. This is because such problems are more likely to appear in the interaction between the user, the data, and the application.

3 PGNIDS Design

As shown Fig.1 PGNIDS consists of DCM(Data Collection Module), ADGM(Audit Data Generation Module), IDPGM(Intrusion Detection Pattern Generation Module), and PGGM(Pattern Graph Generation Module). DCM collects all the incoming packets in Network. ADGM generates the audit data that decides an intrusion and classifies them according to behavior characteristics. IDPGM generates patterns with the classified audit data. PGGM generates pattern graphs with the patterns that IDPGM generates and decides whether it is an intrusion with them.

3.1 Network-Based DCM

PGNIDS in this paper uses libpcap which is called packet capture library provided in LINUX. libpcap captures and filters packets. The necessary contents from the filtered packets are extracted according to packet filtering rule in order to generate only audit data that PGNIDS requires.

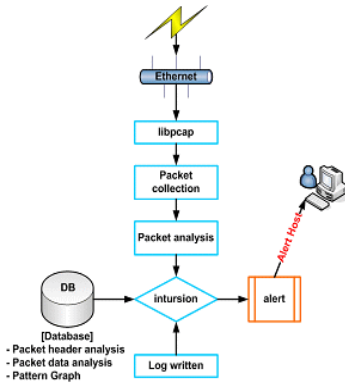


Fig. 1. PGNIDS design

```

struct pcap {
    int snapshot;
    int linktype;
    int tzoff;
    int offset;

    struct pcap_sf sf;
    struct pcap_md md;
    /* Read buffer. */
    int bufsize;
    u_char *buffer;
    u_char *bp;
    int cc;
    /* Place holder for
    pcap_next().
    */
    u_char *pkt;
    /* Placeholder for filter
    code if bpf not in kernel. */
    struct bpf_program fcode;
    char
    errbuf[PCAP_ERRBUF_SIZE];
}

typedef struct pcap pcap_t;
    
```

Fig. 2. The algorithm for capturing packets using libpcap

3.1.1 Packet Capture

Packet capture confirms the contents of all the incoming packets in network. This can apply to various types such as monitoring, network debugging and sniffing for statistics and security for network use. The method capturing packets by using libpcap is as follows. First, the device(NIC) or the file which will be captured is opened, the packets are analyzed, and the device or the file is closed. Libpcap provides various interfaces according to its function. Fig .2 shows the algorithm for capturing packets using libpcap.

3.1.2 Packet Filtering Rule

The rule of packet filter uses source address, source port number, destination address, destination port number, protocol flag, and activity(pass/reject). With these fields, sequential ACL(Access Control List) for filtering packets has to be written. Screening router is the software that decides activity, that is, pass or reject in ACL sequentially. The filtering rule consists of one or several primitives and its format is shown in table. 1.

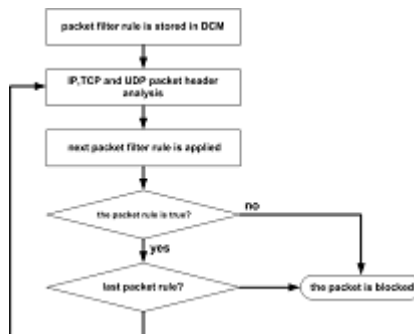
Table 1. The format of packet filter rule

<pre> action [direction] [log] [quick] [on interface] [af] [proto protocol] \ [from src_addr [port src_port]] [to dst_addr [port dst_port]] \ [flag tcp_flags] [state] </pre>

- Action. The action to be taken for matching packets, either pass or block. The default reaction may be overridden by specifying either block drop or block return.
- Direction. The direction the packet is moving on an interface, either in or out.
- Log. Specifies that the packet should be logged via pflogd. If the rule specifies the keep state, modulate state, or syn proxy state option, then only the packet which establishes the state is logged. To log all packets regardless, use log-all.
- Quick. If a packet matches a rule specifying quick, then that rule is considered the last matching rule and the specified action is taken.
- Interface. The name or group of the network interface that the packet is moving through. An interface group is specified as the name of the interface but without the integer appended.
- Af. The address family of the packet, either inet for IPv4 or inet for IPv6.
- Protocol. The layer 4 protocol of the packet. : tcp, udp, icmp, icmp6, a valid protocol name from /etc/protocols, a protocol number between 0 and 255, a set of protocols using a list.
- src_addr, dst_addr. The source/destination address in the IP header.
- src_port, dst_port. The source/destination port in the layer 4 packet header.
- tcp_flags. Specifies the flags that must be set in the TCP header when using proto tcp. Flags are specified as flags check/mask.
- State. Specifies whether state information is kept on packets matching this rule.

3.1.3 Packet Filtering Flow Chart

Packet filtering rules are stored in a particular order and is applied to the packets in that order. Fig.3 shows the flow of packet filtering.

**Fig. 3.** Packet filter operational order

3.2 ADGM(Audit Data Generation Module)

As it is not proper to use the collected data for audit data, ADGM(Audit Data Generation Module) is used to extract only audit data that can decide an intrusion from the collected packets. First, ethernet frame in packet filtering must be divided into IP or ARP packet separately in ethernet layer. In IP layer, IP packet is divided into ICMP, TCP, or UDP separately. That is, in this paper ADGM classifies IP packet into TCP, UDP and, ICMP, generates the audit data and stores them in the database to detect an intrusion. Fig .4 passes the packets of unsigned char type to the pointer of ethernet header. If the packet is IP protocol, Fig .4 shows that it is transferred to the pointer of IP header and is classified into TCP, UDP, and ICMP.

```

void packet_analysis(unsigned char *user, const struct pcap_pkthdr *h,
const unsigned char *p)
{
unsigned int length = h->len;
struct ether_header *ep;
unsigned short ether_type;
length -= sizeof(struct ether_header);
ep = (struct ether_header *)p;
p = += sizeof(struct ether_header)
...
if (ip->protocol == IPPROTO_TCP) { //TCP protocol
tcp = (struct tcphdr *) (P + (iph->ihl * 4) + (tcph->doff * 4));
tcpdata = (unsigned char *) (p + (iph->ihl * 4) + (tcph->doff * 4));
}
...
if (ip->protocol == IPPROTO_UDP) { //UDP protocol
udph = (struct udphdr *) (p + iph->ihl * 4);
udpdata = (unsigned char *) (p_iph->ihl * 4) + 8;
}
...
if (ip->protocol == IPPROTO_ICMP) { //ICMP protocol
icmp = (struct icmp *) ([+iph->ihl * 4]);
icmpdata = (unsigned char *) (p + iph->ihl * 4) + 8;
...
}

```

Fig. 4. Packet Analysis Algorithm

3.3 IDPGM(Intrusion Detection Pattern Generation Module)

The Intrusion detection pattern proposed in this paper is based on that of Snort.

3.3.1 Snort

Snort is a lightweight network IDS that real time traffic analysis and packet logging can be done on the IP network. In addition, it is network sniffer based on libpcap.

That is, it is a tool which monitors, writes, and alarms network traffic that matches intrusion detection rule.

Snort can do protocol analysis, contents detection and the matching and detect the various attacks and the scans such as overflow, stealth port scan, CGI attack, SMB detection.

3.3.2 IDPGM

Intrusion detection pattern is based on that of Snort and Snort consists of the packets of TCP, UDP and ICMP, etc. These packets generates pattern format with backdoor.rules, ddos.rules, dns.rules, dos.rules, exploit.rules, finger.rules, ftp.rules, icmp.rules, info.rules, misc.rules, netbios.rules, policy.rules, rpc.rules, rservices.rules, scan.rules, smtp.rules, sql.rules, telnet.rules, virus.rules, and web-cgi.rules etc. These pattern format is shown in table. 2.

Table 2. Pattern format

```
alert tcp $EXTERNAL_NET any ->
$HOME_NET 80 (content:"|90C8 C0FF FFFF|bin/sh" msg:
"IMAP buffer overflow!");
```

The generated pattern proposed in this paper is stored in a database mysql in order to be used for intrusion detection. Fig .5 shows the structure of TCP data.

```
{
  unsigned short s_port; // source port number
  unsigned short d_port; // destination port number
  int flat; // TCP flag
  char content[200]; // data
  int c_size; // content length of content
  int p_size; // size of payload
  char msg[200]; // message
}
```

Fig. 5. Structure of TCP data

```
struct pattern_graph
{
  char p_name[50] // name of pattern graph
  unsigned short n_id; // node number
  unsigned short s_port; // source port number
  char content[200]; // data
  char msg[200]; // message
}
```

Fig. 6. Structure of Pattern Graph

3.4 PGGM(Pattern Graph Generation Module)

The patterns generated by IDPGM are stored in a database. PGGM generates pattern graphs by analyzing the relationship between patterns, stores them in the database, and prevents a devious intrusion by using the generated pattern graph.

3.4.1 Pattern Detection Algorithm

The Proposed PGNIDS changes AS(Attack Specification Language) to the pattern that is the data structure suitable for PGNIDS to process attacks. The pattern graph is to draw the process of Scenario in tree type and the last node of the tree has no transmission event. Each node of pattern graph means a message. Table. 3 shows the type of message.

Table 3. Message Format

type	<node ID, timestamp, attribute price>
node ID	Each node number in pattern graph
timestamp	Event occurrence hour
attribute price	Attribute price of event



Fig. 7. Pattern graph generation flow chart

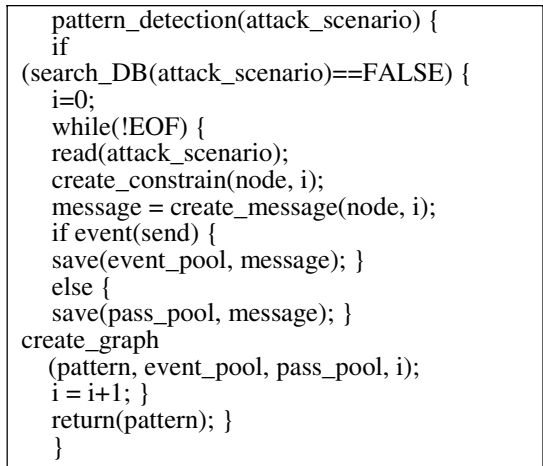


Fig. 8. The algorithm detecting pattern

The message of each node, when the event of the node is transmitted, is stored in the event pool and the rest of events are stored in the pass pool. In the course of changing to pattern graph, some limitations about each node happens. At this time the conditions of the limitations consists of the static limitation condition that has a constant value and the dynamic limitation condition that has a variable value. Fig. 7 shows the flow chart of pattern graph. Fig. 8 shows the algorithm detecting pattern.

When pattern detection algorithm is applied by using the attack scenario of Fig. 9, the generated pattern graph is shown in Fig. 10.

```

ATTACK "sample" [a, b, c]
a {
send(c) : e1[$x=a0;] }
b {
e2[$y = a1;]
send(c) : e3[a2 == $x;] }
c {
e4[]
e5[a3 == $y;]}
    
```

Fig. 9. Attack scenario

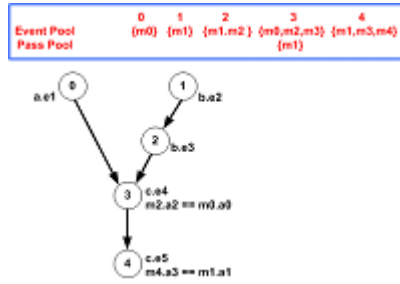


Fig. 10. Pattern graph

4 PGNIDS Simulation

The objects of PGINDS proposed in this paper are all the packets on the network and it is possible for PGINDS to capture packets under TCP, UDP, and ICMP environment by using packet filter. As PGNIDS performs real-time network, it reduces the collection of packets and analysis time and reacts to abnormal attack real-time.

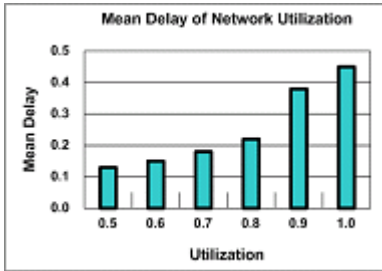


Fig. 11. Network utilization analysis

Table 4. The number of event according data type

Event Type	24 hours
alert	32,448
Log	33,127

4.1 The Analysis of Network Utilization

Fig. 11, when PGNIDS reports the detection information to a server, shows the mean delay time of each event. That is, Fig. 11 shows the delay of transmission packet according to the change of network utilization and that its delay is increasing rapidly over utilization 0.8.

4.2 The Analysis According to Event Type

Table. 4 shows the analysis according to event type processed by packet filtering rule for 24 hours in PGNIDS. Fig. 12 shows the delay of alert and log event. The delay about each event shows the increasing trend according to the network utilization.

4.3 The Analysis Using Pattern Graph

PGNIDS generates pattern graph by using each event, stores it in a database, and detects a devious attack by making pattern graph with the relationship between each event. Fig. 13 shows the detection ration of devious attack according to network delay ratio by using pattern graph between each event.

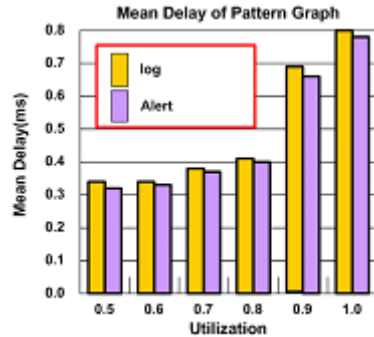
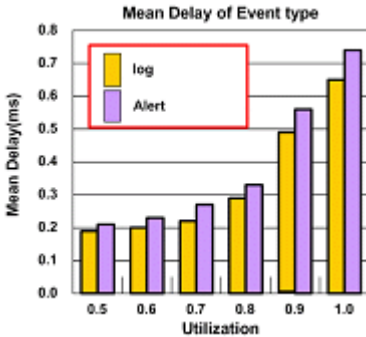


Fig. 12. The analysis according to event type **Fig. 13.** The analysis with a pattern graph

5 Conclusion

PGNIDS is the system that decides an intrusion by collecting network packets real time. An existing IDS, when it estimates an intrusion by reading all the incoming packets in network, takes more time than the proposed PGNIDS does. As this proposed PGNIDS not only classifies the audit data into alert and log through ADGM and stores them in the database, but also estimates the intrusion by using pattern graph that classifies IDPM and event type, Therefore, it takes less time to collect packets and analyze them than the existing IDS, and reacts to abnormal intrusion real time. In addition, it is possible for this to detect the devious intrusion detection by generating pattern graph.

References

1. Byung-Kwan Lee, Eun-Hee Jeong, "Internet security", Namdoo Books, 2005
2. LBNL's Network Research Group
3. <http://www.linux.co.kr/>
4. Kwang-Min Noh, It uses pcap library from linux and packets it catches and it sees v0.3, 2000.09.14, Linux Korean alphabet document project
5. <http://www.snort.org>
6. <http://www.silicondefense.com/snortsnarf>
7. <http://my.dreamwiz.com/winmil/security/snort.htm>
8. <http://www.whitehats.com/>

9. Tsutomu Tone, "1% network principal which decides a success and the failure", Sungandang, 2004
10. <http://www.windowsecurity.com>
11. Dai-il Yang, Seng-Jea Lee "Information security surveying and actual training", Hanbit Media, 2003
12. Rebecca Bace, Peter Mell, NIST Special Publication on Intrusion Detection Systems
13. <http://www.openbsd.org/faq/pf/filter.html>

Experiments and Hardware Countermeasures on Power Analysis Attacks

ManKi Ahn¹ and HoonJae Lee²

¹ Defense Agency for Technology and Quality Assurance,
Kyungpook National University,
Daegu, 706-020, Korea
mkahn@dqaa.mil.kr

² Dongseo University, Busan, 617-716, Korea
hjlee@dongseo.ac.kr

Abstract. Security is a concern in the design of smartcards. It is possible to leak much side channel information related to secret key when cryptographic algorithm runs on smartcards. Power analysis attacks are a very strong cryptanalysis by monitoring and analyzing power consumption traces. In this paper, we experiment Exclusive OR operation. We also analyze the tendency of state-of-the-art regarding hardware countermeasures and experiments of Hamming-Weights on power attacks. It can be useful to evaluate a cryptosystem related with hardware security technology.

Keywords: Side Channel Attacks, Power Analysis, SPA/DPA, Countermeasure, SmartCard.

1 Introduction

The power consumption of a cryptographic device such as smartcard may provide much information about the operations that take place and the involved parameters. In 1999, P.Kocher introduced the so-called side channel attacks based on *simple power analysis* (SPA) and *differential power analysis* (DPA) to recover the secret key[1]. A smartcard, based on the idea of embedding an integrated circuit chip within a ubiquitous plastic card, can execute cryptographic operations and provide high reliability and security. Recently, however, this had been a target of the side channel attacks.

This paper¹ analyzes the tendency of state-of-the-art regarding hardware countermeasures and experiments of Hamming-Weights on power attacks, and experiments Exclusive OR operation in smartcards. It will be discussed in detail in section 3. The remainder of this paper is organized as follows: Section 2 overviews power attacks, while section 3, We experiment on power analysis attacks. Section 4 analyzes state-of-the-art regarding hardware countermeasures. Conclusion is presented in section 5.

¹ This research was supported by University IT Research Center Project.

2 Power Analysis Attacks

The power consumption of hardware circuit is a function of the switching activity at the wires inside it. Since the switching activity is data dependent, it is not surprising that the key used in a cryptographic algorithm can be inferred from the power consumption statistics gathered over a wide range of input data. These attacks have been shown to be very effective in breaking smartcards. These attacks are called power analysis attacks which are non-invasive attacks.

Simple power analysis(SPA) consists of observing the variations in the global power consumption of the chip and retrieving from it some information which can help to identify any secret key or value. A special kind of SPA, the so called Hamming-weight attacks exploit a strong relations between the Hamming-weight and the power consumption trace.

Differential power analysis(DPA) is more sophisticated than the SPA. The attacker identifies some intermediate value in the cryptographic computation that is correlated with the power consumption and dependent on the plaintext and the key. The attacker divides the traces into groups according to the intermediate value predicted by current guess at the key and the traces corresponding plaintext. If the averaged power trace of each group differs noticeably from the other, it is likely that the current key guess is correct. Incorrect key guesses should result in all groups having very similar averaged power traces, since incorrectly predicted groups having very similar averaged power traces.

Recently, there are many open questions regarding reconfigurable hardware devices, such as Field Programmable Gate Arrays(FPGAs), as a module for security functions. The use of FPGAs is highly attractive for a variety of reasons that include algorithm upload or modification, architecture efficiency, and costs. However, FPGAs will be targeted of the one-to-one copy, reverse-engineering, and physical attacks. Therefore, many people discuss and experiment vulnerabilities of modern FPGAs against the threat[2][3][4][5][6][7][8][9]. They used either a microchip PIC 16F84A microcontroller, ATMEL AT89S8252, a Xilinx XCV800, Virtex-E FPGA, or ARM CM7TDMI core and used MATLAB, C-programs as statistical analysis tool etc.

A PINPAS(Program Inferred Power Analysis in Software) tool supports the testing of algorithms for vulnerability to SPA/DPA. The tool is especially useful as an aid in the design of both cards(hardware) and algorithms(software)[10][11].

The masking method is the usage of masked logic. However, that does not prevent DPA attacks, because Glitches occur in every CMOS circuit. The Glitches are that the transitions at the output of a gate that occur before the gate switches to the correct output[12].

3 Experiments of Power Attacks on Smartcard

3.1 Experiments of Hamming-Weights

Now, we will carry out the experiments of Hamming-Weights[1] using data transition in smartcard. The instruction takes the Exclusive OR operation(XOR) of

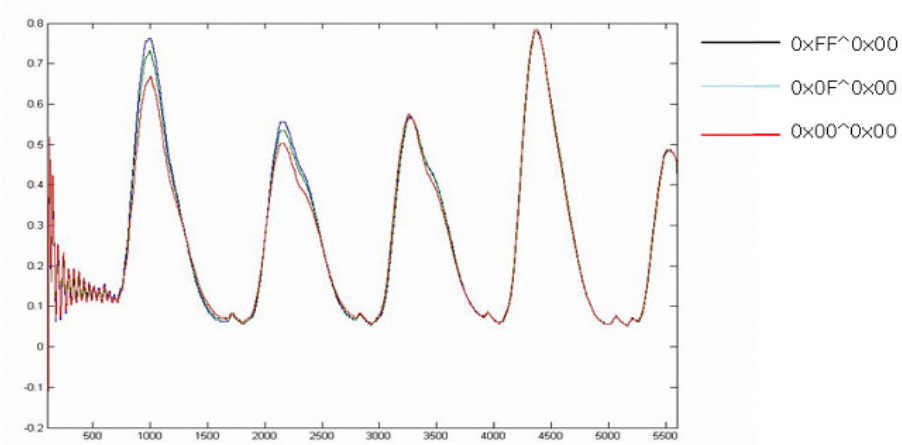


Fig. 1. Power traces of several XOR operations over 1,000 traces

two 8-bit values. The experimental results are shown in figure 1. As the below results, the plot confirms the assumption about the measurability of Hamming-Weights leakage. we need approximately 1,000 measurements to identify the correct plot.

3.2 Experiments of DPA

The plaintexts are prepared that only the data at the output of the 1st S-Box would be different in the first round of block cipher. Further details of the S-Boxes are omitted, but it handles the main ingredients of an algorithm like block ciphers(DES,AES). The smartcard is assumed to leak information about secret values transported on the memory bus. The potential power source for SPA/DPA is the value of a operand XOR secret key which can be calculated from the known operand and a guessed secret key.

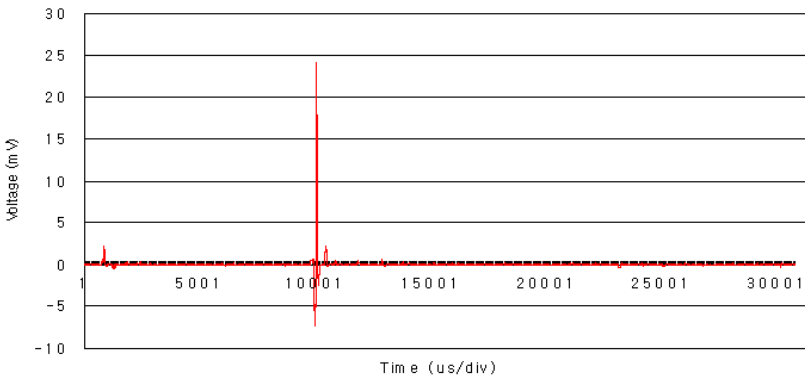


Fig. 2. The differential power traces for the correct key guess

In the criterion, we generated power traces and be split into two groups with Hamming-weights larger and smaller than 4.

By performing several XOR operations with S-Boxes, A difference trace was obtained by subtracting the average traces for each of the two groups. We gather approximately 5,000 measurements. Figure 2 show that the correlation could be observed.

4 Hardware Countermeasures on Power Analysis Attacks

The advantages of software implementations are the ease of use, the ease of upgrade, the portability, low development costs, low unit price and flexibility. Software implementations offers moderate speed, slow the execution process compared to hardware system. Hardware implementations are more secure because they cannot as easily be read or modified by an attackers as software. Hardware countermeasures offer deal either with some form of power trace smoothing or with transistor-level changes of the logic[4]. The goal of countermeasures against DPA attacks is to completely remove or at least to reduce this correlation, i.e. the addition of noise with noise-generators of the filtering of the power traces[13], the insertion of random delays[14], the use of capacitor or dummy bus, internal clock generator including random clock jittering, static complementary CMOS logic[15], or the usage of masked logic, but that does not prevent DPA attacks, because of Glithes occur in every CMOS circuit[12].

4.1 Countermeasures of Logic Level

We summarize security problems produced by attacks against hardware implementations. To be resistant against the SPA/DPA, various countermeasures have already been proposed. The protection against power analysis attacks involved implementing hardware based on a power attacks resistant logic with constant power consumption[16]. It depends on both the values and transitions, i.e. the Hamming-weights between consecutive data values, yet this is quite expensive to implement. Therefore, we analyze another power attack resistant hardware-type and state-of-the-art skill.

Dual-rail method is to render information about Hamming-weights of secret values completely useless, dual-rail logic provide attackers with the meaningless Hamming-weights of values, because these values are always the same. An implementation of this method in hardware can be efficient and transparent to the algorithm running on smartcard. This method used precharge logic. Every signal transition is represented with a switching event, in which the logic gate charges a capacitance. But at a price, the hardware resources have to be doubled in size[10]. Dual-rail encoding can be similarly used to pass data and an alarm signal by using the 11 value to indicate an alarm (00 is used to pass a clear signal; 01 and 10 representing logical-0 and logical-1 respectively). Asynchronous logic(the self-timed circuits) can be made far less susceptible to power attacks, simply slowing down when the supply voltage dips rather than malfunctioning. By contrast, the

self-timed circuits are consumed considerable silicon area(nearly three times the area of the synchronous one) and slower than the synchronous one[17][18][19].

A dynamic and differential CMOS logic is presented in which a gate always uses a fixed amount of power. Sense Amplifier Based Logic (SABL)[16] uses advanced circuit techniques to guarantee that the load capacitance has a constant value. SABL completely controls the portion of the load capacitance that is due to the logic gate. The intrinsic capacitances at the differential in and output signals are symmetric and additionally it discharges and charges the sum of all the internal node capacitances. A major disadvantage is the non-recurrent engineering costs of a custom designed cell library development. SABL also suffers from a large clock load, as is common to all clocked dynamic logic styles and uses two times the area and power of other CMOS logic.

4.2 Countermeasures of Operation Level

Secure instruction based on a pipeline architecture execute sequences of instruction(i.e. fetch, decode, execute, write). This is implemented by the electronics of the microcontroller rather than by software addition. However, this countermeasure is only implemented with RISC(Reduced Instruction Set Computer) architecture in which the instructions are read and executed in parallel. RISC architectures using a so called "pipeline" method make it possible to interleave several instructions by several instructions in the same clock cycle. Therefore, the waiting time is introduced randomly between the sequences of instruction. In other words, there is instruction set architecture of pipelined smart card processor with secure instructions to mask the power differences due to key-related data[20].

5 Conclusion

We have experiments of Hamming-Weights using Exclusive OR operation(XOR) on power attacks. Experimental results have demonstrated that the instruction with the different value of Hamming-Weights can make different power traces. Therefore, at the part of hardware countermeasures, A logic designer must consider DPA-resistant CMOS logic in smartcard. Besides, we also analyze the tendency of state-of-the-art regarding hardware countermeasures. Side-channel resistance cannot be isolated at one abstraction level. It can be useful to evaluate a cryptosystem related with hardware security technology.

References

1. P. Kocher, J. Jaffe, and B. Jun, "Differential Power Analysis," *In Proceedings of Advances in Cryptology-CRYPTO '99*, LNCS 1666, pp. 388-397, Springer-Verlag, 1999.
2. Larry T. MaDaniel III, "An Investigation of Differential Power Analysis Attacks on FPGA-based encryption Systems", available to scholar.lib.vt.edu, *Master of Science in Electrical Engineering*, May, 2003.
3. Siddika Berna Ors, Elisabeth Oswald and Bart Preneel, "Power-Analysis Attacks on an FPGA-First Experimental Results", *In Proceedings of CHES 2003*, LNCS 2779, Springer-Verlag, pp. 35-50. 2003

4. Thomas Wollinger and Christof Paar, "How Secure Are FPGAs in Cryptographic Applications(Long version)", Report 2003 /119, IACR, 2003. available on <http://eprint.iacr.org>
5. Chin Chi Tiu, "A New Frequency-Based Side Channel Attack for Embedded Systems", *A Master thesis, in the University of Waterloo*, 2005
6. Ryan Junea, "POWER ANALYSIS ATTACKS :: A Weakness in Cryptographic Smart Cards and Microprocessors", *Bachelor of Computer Engineering & Bachelor of Commerce*, November, 2002
7. Elisabeth Oswald, "On Side-Channel Attacks and the Application of Algorithmic Countermeasures", *A PhD Thesis in Graz University of Technology*, IAIK, May, 2003
8. Stefan Mangard, " Calculation and simulation of the Susceptibility of Cryptographic Devices to Power-Analysis Attacks", *A Diploma Thesis, in Graz University of Technology*, IAIK, 2003
9. KULRD & SCARD Consortium, "Side Channel Analysis Resistant Design Flow", IST-2002-507270, SCARD-KULRD-D4.1, 2005, available on <http://www.scard-project.org>.
10. J. den Hartog and others, "PINPAS : a tool for power analysis of smartcards", in *SEC 2003, IFIP WG 11.2 Small Systems Security*, pp. 447-451, 2003
11. J.I den hartog, and E.P. de Vink, " Virtual Analysis and Reduction of Side-Channel Vulnerabilities of Smartcards", available on <http://www.win.tue.nl/ecss>, 2005.
12. Stefan Mangard, Thomas Popp, and Berndt M. Gammel, "Side-Channel Leakage of Masked CMOS Gates", *Topics in Cryptology - CT-RSA2005*, LNCS 3376, pp. 351-365, Springer-Verlag, 2005.
13. Kris Tiri and Ingrid Verbauwhede. "A Logic Level Design Methodology for a Secure DPA Resistant ASIC or FPGA Implementation.", In *DATE 2004*, pp. 246-251. IEEE Computer Society, 2004.
14. Stefan Mangard. "Hardware Countermeasures against DPA. A Statistical Analysis of Their effectiveness. In *proceedings of Cryptology-CT-RSA 2004*, LNCS 2964, pp. 222-235. Springer-Verlag, 2004.
15. Kris Tiri and Ingrid Verbauwhede, "A VLSI Design Flow for Secure Side-Channel Attack Resistant ICs" In *Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE05)*, 2005.
16. Kris T., Moonmoon A., and Ingrid V., "A Dynamic and Differential CMOS Logic with Signal Independent Power Consumption to withstand Differential power Analysis on Smart Cards" In *28th European Solid-State Circuits Conference*, 2002.
17. K. J. Kulikowski, Ming Su, A.Smirnov, A. Taubin, M. G. Karpovsky, and Daniel M., "Delay Insensitive Encoding and Power Analysis: A Balancing Act", In *11th IEEE International Symposium on Asynchronous Circuits and Systems: ASYNC'05*, pp. 116-125, 2005.
18. S. Moore and others, "Improving SmartCard Security using Self-timed Circuits", available on http://actes.sstic.org/SSTIC03/Rump_sessions, 2003.
19. Simon Moore, Ross Anderson, Robert Mullins and George Taylor, "Balanced self-checking asynchronous logic for smart card applications" in *the Microprocessors and Microsystems Journal*, 2003
20. Feyt, "Countermeasure method for a microcontroller based on a pipeline architecture", *US PATENT 20030115478 A1*, 2003
21. Elisabeth Oswald, Stefan Mangard, Norbert Pramastaller, Vincent Rijmen, "A Side-Channel Analysis Resistant Description of the AES S-Box.", *FSE 2005, Revised Selected Papers*, LNCS 3557, pp. 413-423, Springer-Verlag , 2005.
22. Kris Tir and Ingrid Verbauwhede, "Simulation Models for Side-Channel Information Leaks" *ACM 1-59593-058-2/05/0006, DAC 2005*

Information System Modeling for Analysis of Propagation Effects and Levels of Damage

InJung Kim¹, YoonJung Chung², YoungGyo Lee¹, Eul Gyu Im³,
and Dongho Won^{1,*,**}

¹Information Security Group, School of Information and Communication Engineering,
Sungkyunkwan University

cipher@etri.re.kr, {yglee, dhwon}@security.re.kr

²Electronics and Telecommunications and Research Institute

yjjung@etri.re.kr

³College of Information and Communications, Hanyang University

imeg@hanyang.ac.kr

Abstract. The number of newly developed information systems has grown considerably in their areas of application, and their concomitant threats of intrusions for the systems over the Internet have increased, too. To reduce the possibilities of such threats, studies on security risk analysis in the field of information security technology have been actively conducted. However, it is very difficult to analyze actual causes of damage or to establish safeguards when intrusions on systems take place within the structure of different assets and complicated networks. Therefore, it is essential that comprehensive preventive measures against intrusions are established in advance through security risk analysis. Vulnerabilities and threats are increasing continuously, while safeguards against these risks are generally only realized some time after damage through an intrusion has occurred. Therefore, it is vital that the propagation effects and levels of damage are analyzed using real-time comprehensive methods in order to predict damage in advance and minimize the extent of the damage. For this reason we propose a modeling technique for information systems by making use of SPICE and Petri-Net, and methods for analyzing the propagation effects and levels of damage based on the epidemic model.

Keywords: Risk analysis, Intrusion, Damage propagation, Safeguard, Epidemic.

1 Introduction

Security risk analysis [1] of information systems is the best means of eliminating vulnerabilities from information security services and safely controlling the systems against potential threats. Currently, information systems operate in various environments with extended areas, a large number of assets and interoperations with heterogeneous systems such as controlling systems. This situation has enabled risk analysis

* Corresponding author.

** This work was supported by the University IT Research Center Project funded by the Korean Ministry of Information and Communication.

simulation of information systems to emerge as a field of keen interests and to bring innumerable studies and discussions. An important prerequisite to perform simulations is to create an environment in which analysis of the propagation effects and levels of damage to information systems can be analyzed. In such a simulated environment, an analysis of the activities on information systems, resources and information flow should be conducted in order to evaluate the affects of cyber intrusions on the information systems. For the analysis of activities on information systems, we model information systems through the SPICE model [3] and Petri-Net [4] for circuit design, and analyze propagation effects and levels of damage by applying the epidemic model [2][24]. The epidemic model has been studied for the propagation of worms; we will use the model to analyze all cyber intrusions as well as worms.

It is normally difficult to identify which intrusion causes damage. Once an intrusion takes place, the related functions of the information system are degraded, or the intrusion shuts down some of information systems. After recognizing the symptoms, system administrators will begin to establish safeguards for the damage. Once these safeguards have been established, recovery procedures for the affected systems may begin. Meanwhile, damage from the intrusions might have been propagated to other systems via unspecified routes, and the scope of the damage increases accordingly. In such a case, damage continues to occur until safeguards are established to prevent future intrusions.

Therefore, we propose a modeling mechanism to assist the analysis of possible intrusions in advance. Our proposed modeling mechanism will help system administrators to analyze cyber threats and establish effective safeguards for prevention and recovery from intrusions.

2 Related Work

2.1 Information System Modeling

In the most organizations, information systems are modeled to show network configurations simply using Microsoft PowerPoint or VISIO. This type of modeling is capable of showing the current status and connection features of assets only; thus, it is difficult to analyze damage propagation using this kind of modeling, since the modeling is not capable of showing job flows or predicting the propagation effects of damage occurred. Flow charts or state transition diagrams can be used to identify information flow, but these approaches have limitations to analyze and identify threats from the overall network configurations of information systems. More recently, the state transition diagrams [5] have been extended for direct representation of sequence and elements of events as well as simple illustration of behaviors and results of cyber intrusions in the systems through Deterministic Finite State Machine (DFSM) [6] or Colored Petri-Net [7]. The state transition diagram approach configures only the effects and damage routes of cyber intrusions, so it has some limitations in risk analysis: This approach is not capable of incorporating the unique features of respective assets when representing the information system as a model, and this approach illustrates the distribution of damage unevenly according the directions of propagation. To overcome these shortcomings, the SPICE model has been introduced to analyze transient effects.

2.2 SPICE Modeling

The Simulation Program with Integration Circuit Emphasis, or SPICE [8], is a program to simulate simple electronic circuits based on equivalent circuits for the respective elements. With SPICE, users can design, edit, and simulate electronic circuits, and users can compile characteristics of elements and circuit configurations in a library for later analysis. However, the SPICE model may contain excessive unnecessary information assets to cover each asset in the entire information systems and may contain more complex designs than network layouts. This may cause difficulties in analysis of the routes for cyber intrusions and damage incurred by the intrusions. Therefore, a new modeling technique is required for simple and easy analysis of damage routes, so that the modeling technique can be used for risk analysis of information systems.

2.3 Risk Analysis

Many studies on risk analysis of information systems are currently in progress in different fields. The major three domains are as follows:

- Risk analysis processes and risk-level calculation
- Design and development of risk analysis tools
- Studies on control items and guidelines

Several risk analysis processes have been developed, including GMITS [9], CSE (Communications Security Establishment) [10], HAZOP (Hazard and Operability study) [11], FTA (Failure Model and Effect Criticality Analysis) [12], OCTAVE [13], and CORAS [14]. In Korea, a process called PRAHA [15] has been developed and utilized for analysis and assessment of vulnerabilities of systems in governmental or public organizations. Risk analysis tools include CRAMM (CCTA Risk Analysis and Management Methodology) [16], BDSS [17], and Buddy System [18]. The BS7799 [19] and IT Baseline Protection Manual from BSI [20] are under study for control items and calculation of criteria. Most of the above processes or tools can be used for risk analysis; however, they are somewhat limited in their analysis of the scope of damage and effects caused by intrusions.

2.4 Intrusion Damage Estimation

The damage calculation proposed in [21] simply calculates the values of damaged assets, labor costs, recovery expenses in a quantitative manner for the duration of intrusions. This methodology does not really contain a technique for real-time analysis of the rapid changes in information systems because of different intrusion accidents; nor is it capable of analyzing the routes and affects of damage incurred. Some studies [22] are currently in progress to analyze the extent of damage using the propagation model for worms by making use of certain epidemic models [2], and estimate the levels of damage through real-time analysis of the availability of information systems. However, no study on overall damage propagation has yet been completed.

3 Information System Modeling

The modeling of information systems is a must for the analysis of the propagation effect and level of damage caused by intrusion. To model systems, it is necessary for the target systems and assets to be defined, while the functional restrictions of systems such as objectives and configuration shall be explicitly specified. However, the authors of the study illustrate the mutual reliance between the assets comprising a system and the similar functions of the assets in a block diagram [23]. To do this, the authors illustrate system modeling as shown on Fig. 1, define the elements as follows:

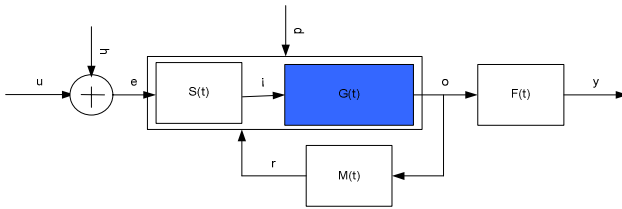


Fig. 1. Structure of information system

Info-Infra Model $IM = \langle G(t), S(t), M(t), F(t), e, r, y, t \rangle$

- $G(t)$: Information system
- $S(t)$: Information security system or encryption system
- $M(t)$: Monitoring or control system
- $F(t)$: Communication system or security guard for output data
- e : Input into information security system
- r : Input from monitoring or control system
- y : Output of information system or input of linked system
- h : Hacking from outside of system
- d : In-house intrusion accident
- u : Control level of input of, or access from, users
- t : Time

Where, $\{G_i\}$ is elementary assets including servers, networks and PCs.

The modeling of information systems for risk analysis is configured in a block diagram. As shown in Fig. 2, it is assumed that a web server, an application server and a database server reside inside an information system. Users are allowed to access the Internet while they perform e-library jobs. However, there are difficulties in analyzing the propagation effects and levels of damage caused by cyber intrusions in information systems with the block diagram of the information system. Representing the information system using modeling as shown in Fig. 3 may cause hacking threats from the Internet; therefore, it should be possible to easily recognize in-house intrusion accidents and clearly acknowledge target assets protected by the information system. Therefore, this simple network block diagram enables analysis of the propagation effects and levels of damage.

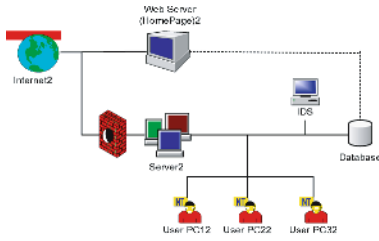


Fig. 2. Block diagram of a common information system

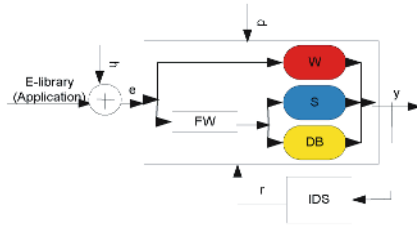


Fig. 3. Configuration of information system through modeling

4 Analysis of Levels of Damage

To analyze the propagation effects of damage using the suggested modeling, each asset and network features should be configured in the modeling. For this purpose, we define levels of damage to information systems as follows:

The fault conditions and the range of assets subject to damage should be identified in order to analyze the entire level of damage to an information system. The current epidemic model is as shown in Fig. 4:

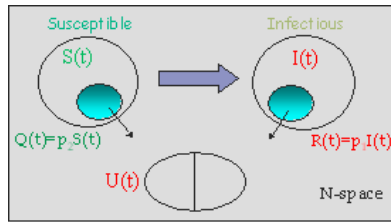


Fig. 4. Epidemic model

The epidemic model allows analysis of the infection feasibility of susceptible assets from those infected by worms, and calculation of the number of infected assets based upon the results of the analysis. Assets can be removed after infections. The equation is as follows:

$$\begin{aligned}
 I'(t) &= \beta I(t)S(t) - U'(t) \\
 U(t) &= Q(t) \cup R(t) = \{p_2 \cap S(t)\} \cup \{p_1 \cap I(t)\} \\
 S(t) \cup I(t) \cup U(t) &= N
 \end{aligned}$$

$S(t)$: susceptible Assets, $I(t)$: Infectious Assets,
 $U(t)$: removed Assets from Infection, N : Total Asset Number,
 β : Infection Ability, p_1, p_2 Recovered Ability

This model does not suggest the propagation effect and level of damage occurred during a period of activities such as analyzing the causes of intrusion, establishing

safeguards, and recovering the systems from damage. Furthermore, this model is applicable to worms only, and is not capable of identifying the damage probabilities of different intrusions or the levels of damage to entire systems. Therefore, additional elements should be defined to expand the model to other cases. We expanded the epidemic model to include the following factors for intrusion:

1. Levels of damage by judging if damage is caused by normal operations or by intrusions when the assets are loaded
2. A scope and levels of asset infections when an intrusion takes place
3. Levels of protection in phases of intrusion elimination after the causes of intrusions have been identified

Each asset will be infected if it is susceptible to intrusions. In such cases, the level of damage to each asset is calculated using a relational function with the uncertainty of asset infections. The level of damage to the entire information system is calculated using functions relevant to existing safeguards and the level of recovery attained through the safeguards. An information system is the sum of its assets: each asset faces unique threats and it is vulnerable and subject to damage due to these threats and its vulnerabilities. The level of damage of a system over time can be represented in a function. The final results of the analysis are as follows when the probability and the uncertainty of infection are included.

$$\begin{aligned} \{I_i(t)\} &= \{P_i(t)\} \cap \{X_i(t)\} \\ \{R_i(t+d)\} &= f_{R1}[\{I_i(t+d)\}, \{\epsilon_i(t+d)\}] + f_{R2}[\{I_i(d)\}, \{\epsilon_i(d)\}] \\ \{G_i(t)\} &= f_T[\{A_i(t)\}, \{Z_i(t)\}, f_R[\{P_i(t) \cap X_i(t)\}, \{\epsilon_i(t)\}]] \\ &= f_R[\{\rho_i(t) \cap \{A_i(t)\}\}, \{\sigma_i \cap \{Z_i(t)\}\}, f_R[\{P_i(t) \cap X_i(t)\}, \{\epsilon_i(t)\}]] \end{aligned}$$

$$M(t) = G(t)/N \times 100$$

$\{P\}$: Intrusion, $\{X\}$: Weaknesses attributed to intrusion
 $\{R\}$: Level of damage to infected asset, $\{\epsilon\}$: Uncertainty of infection,
 $\{A\}$: Level of existing safeguards, $\{d\}$: Time delay in analyzing infection,
 $\{R\}$: Level of future safeguards, ρ, σ : Probability of infection
 M : Level of damage (%)

5 Damage Propagation and Calculation Using Modeling

We have performed the following case study for the analysis of security risks using the suggested modeling. The test environment for the case study was configured as shown in Fig. 5, and modeling was performed as shown in Fig. 6. The in-house network is employed for business management, and the home page is operated in the outside network. The data protection system is installed and operated on each client. The switches and hubs at both ends to build the information security system are removed from the major asset list, since they are not regarded as major assets. The systems operating in a duplex structure and a dual configuration are indicated as overlapping, since they are identical in terms of the probabilities of threats and vulnerabilities.

Configuring the information system in the manner illustrated above allows us to analyze the propagation effects and levels of damage as well as assets, threats, vulnerabilities, and safeguards.

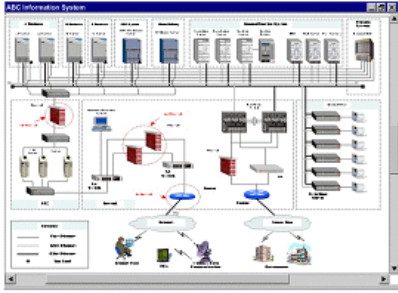


Fig. 5. Configuration diagram of a information system

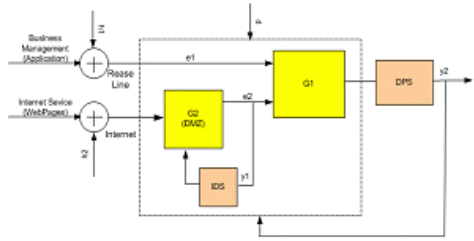


Fig. 6. Modeling of a information system

5.1 Structure of Assets

The structure of assets is as follows: threats and vulnerabilities regarding an intrusion in question are distinguished, and the probability of infections of the distinguished items is defined. Table 1 contains an example of the asset parameters for a Windows 2000 server. As shown in Table 1, the Windows 2000 Server in an information system is subject to threats of improper password control (2.19), an absence of logging policies (2.35), and a deficiency in training programs for operators (3.09). Infection information and a scope, and relational functions are recorded with time intervals. The system is susceptible to buffer overflow attacks and format string attacks, which can cause damage of up to 70% and 40% of the server assets respectively.

Table 1. File structure of the asset Parameters

Table 2. Modeling results for each asset

```
! Windows2000 Server
! PARAMETER DATA
! Threat Table
! Table number: Threat factor (1-5):
  Damage_factor (%)
2.19: 1, 2, 2, 3, 4, 5: linear_function (a)
2.35: 1, 2, 3, 4, 5, 5: exp_function (a)
3.09: 1, 3, 5: log_function (a)
... ..
! Vulnerability Table
! CVE ID: Threat_factor (1-5):
  Damage_factor (%)
CAN-2000-1186: 4: 70
CAN-2003-1022: 3: 40
... ..
! Risk Table
...
! Damage Table
...
! SafeGuard Table
...
```

```
! Information System Net-list
DIM
1 min! Time Interval
IN 0 1 Attack
OUT 0 1 Risk, Damage
! Assets
SERVER 1 2 S(WIN2000, SP2) !S
DB 2 3 D(ORACLE, V8.0) !D
PC 2 4 PC(WINXP, 3) !PC set 3
SERVER 1 3 S(HP, SP3, APACHE) !S
.....
```

5.2 Netlist File

A netlist file is defined as a basic network-structured input file for an information system. When an information system is developed as shown in Fig. 5 with a netlist file as shown in Table 2, details are displayed as follows. Since most of assets are connected with switches or hubs, location determinations of switches and hubs as nodes allow simple development of netlist files.

First of all, identify the assets currently used by the information system, and mark the asset type, and the numbers and the values of adjacent nodes where the assets reside. The core of the netlist file illustrates the information system based on node numbers.

The modeling of the information system as shown in Fig. 5 is displayed as shown in Fig. 7, and simulation is made for the propagation effects and calculations of damage to show the results as in Fig. 8. The results indicate that the level of damage to the information system escalated from 1 to 5 in 120 seconds after the intrusion, with the total damage exceeding 80%. However, operations for the emergency recovery measures make the entire damage level become 2 and reduce the damage to 40%. As described above, the modeling of an information system enables convenient real-time analysis of the scopes and levels of damage.

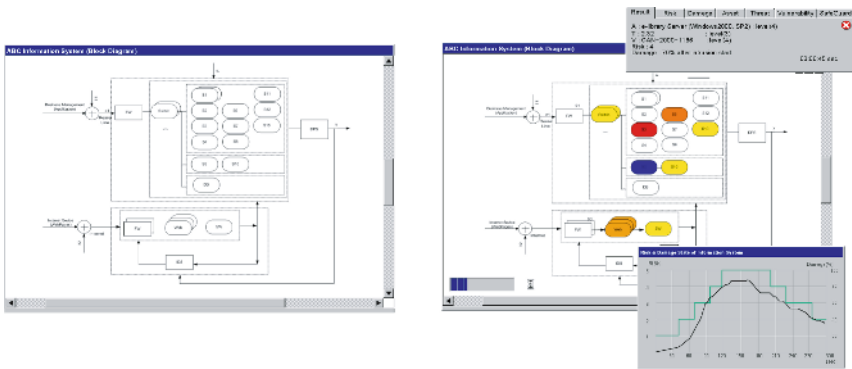


Fig. 7. Modeling of an information system **Fig. 8.** Propagation effects and levels of damage

6 Conclusions

Security risk analysis of information systems is an essential task. However, the current analytic techniques are, in general, of a static nature, and are not capable of illustrating the propagation effects of damage to information systems and the appropriateness of role operations for information security systems. We suggest techniques to analyze the propagation effects and levels of damage while resolving the above problems.

The techniques suggested in this study have been shown to be capable of analyzing the damage flow within information systems and the effects of damage in a given time interval. This means that when an intrusion takes place, the level of infection and its route are analyzed by identifying the threats and vulnerabilities of various types

and levels of intrusions, while the variation in levels of damage and damage propagation effects are analyzed by employing pre- and post-information safeguards. In short, these techniques allow safeguards for information systems to be defined in a relatively short period of time as well as real-time analysis of appropriateness of the safeguards in order to execute more stable and efficient security risk analysis.

References

- [1] Hoh Peter In, Young-Gab Kim, Taek Lee, Chang-Joo Moon, Yoonjung Jung, Injung Kim, "Security Risk Analysis Model for Information Systems," LNCS 3398, Systems Modeling and Simulation: Theory and Applications: Third Asian Simulation Conference, AsianSim 2004.
- [2] Yun-Kai ZHANG, Fang-Wei Wang, Yu-Qing ZHANG, Jain-Feng MA, "Worm Propagation Modeling and Analysis Based on Quarantine," Infosec04, November 14-16, 2004, ACM ISBN:1-58113-955-1.
- [3] Kwang Min Park, Dong Kwang, PSpice Understanding and Application (revised), 1992, ISBN 89-85305-02-6.
- [4] W. Reisig, Petri Nets, An Introduction, EATCS, Monographs on Theoretical Computer Science, W. Brauer, G. Rozenberg, A. Salomaa (Eds.), Springer Verlag, Berlin, 1985.
- [5] Edward Yourdon, Modern Structured Analysis, Prentice-Hall, 1989.
- [6] Paul E. Black, ed, "Deterministic finite state machine", Dictionary of Algorithms and Data Structures, NIST. <http://www.nist.gov/dads/HTML/determFinitStateMach.html>
- [7] L.M. Kristensen, S. Christensen, K. Jensen: The Practitioner's Guide to Coloured Petri Nets. International Journal on Software Tools for Technology Transfer, 2 (1998), Springer Verlag, 98-132.
- [8] Paul Tuinenga, SPICE: A Guide to Circuit Simulation and Analysis Using PSpice (3rd Edition), Prentice-Hall, 1995, ISBN 0-13-158775-7.
- [9] ISO/IEC TR 13335, Information technology - Guidelines for the management of IT Security: GMITS, 1998.
- [10] CSE (Canadian Security Establishment), "A Guide to Security Risk Management for IT Systems", Government of Canada, 1996.
- [11] MacDonald, David/ Mackay, Steve (EDT), Practical Hazops, Trips and Alarms (Paperback), Butterworth-Heinemann, 2004.
- [12] RAC, Fault Tree Analysis Application Guide, 1991.
- [13] CMU, OCTAVE (Operationally Critical Threat, Assets and Vulnerability Evaluation), 2001. 12.
- [14] Theo Dimitrakos, Juan Bicarregui, Ketil Stølen. CORAS - a framework for risk analysis of security critical systems. ERCIM News, number 49, pages 25-26, 2002.
- [15] Young-Hwan Bang, Yoonjung Jung, Injung Kim, Namhoon Lee, Gangsoo Lee, "Design and Development of a Risk Analysis Automatic Tool," ICCSA2004, LNCS 3043, pp.491-499, 2004.
- [16] <http://www.cramm.com>, CRAMM
- [17] Palisade Corporation, @RISK, <http://www.palisade.com>.
- [18] Countermeasures, Inc., The Buddy System, <http://www.buddysystem.net>
- [19] Information Security Management, Part 2. Specification for Information Security Management System, British Standards Institution (BSI).
- [20] BSI, <http://www.bsi.bund.de/english/gshb/manual/index.htm>, 2003.

- [21] Thomas Dubendorfer, Arno Wagner, Bernhard Plattner, "An Economic Damage Model for Large Scale Internet Attacks," Proceedings of the 13th IEEE International Workshops on Enabling Technologies Infrastructure for Collaborative Enterprise (WET ICE'04) 1524-4547/04.
- [22] InJung Kim, YoonJung Chung, YoungGyo Lee, Dongho Won, "A Time-Variant Risk Analysis and Damage Estimation for Large-Scale Network Systems," ICCSA2005, LNCS3043, May 2005.
- [23] Injung Kim, YoonJung Jung, JoongGil Park, Dongho Won, "A Study on Security Risk Modeling over Information and Communication Infrastructure," SAM04, pp. 249-253, 2004.
- [24] M. Liljenstam, D.M. Nicol, V.H. Berk, and R.S. Gray, "Simulating Realistic Network Worm Traffic for Worm Warning System Design and Testing," In Proceedings of the 2003 ACM workshop on Rapid Malcode, pp.24-33, ACM Press. 2003

A Belt-Zone Method for Decreasing Control Messages in Ad Hoc Networks

Youngrag Kim¹, JaeYoun Jung¹, Seunghwan Lee², and Chonggun Kim^{1,*}

¹ Dept. of Computer Engineering, Yeungnam University,
214-1 Dea-dong Gyeongsan-si Gyeongsangbuk-do (712-749 Korea)
yrkim@yumail.ac.kr, cgkim@yu.ac.kr

² SoC R&D Center, System LSI division, Samsung Electronics Co., Ltd
seung1972@hotmail.com

Abstract. MANET(Mobile Ad Hoc Network) is the composite technology of mutual wireless connections of nodes in mobile networks. In AODV(Ad hoc On-demand Distance Vector) Routing, all the nodes have to receive route request messages, and rebroadcast the route request messages for others. It causes a lot of network traffic overheads. In this study, we propose a belt-zone selection method for decreasing the number of RREQ messages. All nodes that receive the RREQ message don't rebroadcast, but only the nodes within a selected zone rebroadcast it for setting a routing path. The belt-zone a logical concentric area is decided by the signal strength from RREQ sender. We also provide an efficient belt-zone selecting sequence by simulations. In high density networks, an intermediate area within several logical concentric areas from the sender must be selected as the primary area. But, in low density networks, the far area from the sender is better to be selected as the primary belt-zone area. By applying the proposed belt-zone mechanism to AODV networks, a lot of control messages can be decreased on various load situations.

1 Introduction

MANET is a network with no fixed infrastructure, such as underground cabling or base stations, where all nodes are capable of moving and can be connected dynamically in an arbitrary manner. Nodes, in MANET, work as routers to discover and maintain routes to other nodes[1]. An important challenge in the design of ad hoc networks is the development of a dynamic routing protocol that can efficiently find a routing path between two communicating nodes. MANET routing protocol consists of table-driven protocol and on-demand protocol[2-5]. In table-driven protocol, each mobile node maintains routes to all nodes in the network. It requires periodic routing advertisements to be broadcasted by each node. In a dynamic mobile network, the topology information is soon out of date, and the propagation of routing information is too slow to be accurate. But the other hand, an on-demand routing protocol creates routes only when the source node has data to transmit to destination. In AODV, which is a representative on-demand routing protocol, a source node broadcasts a route

* Correspondence author.

request control message(RREQ) to discover a routing path to a destination node. Neighbors of the sender node receive it. Each node that has received the RREQ message rebroadcasts it to its own neighbors. This process continues until the RREQ reaches to the destination node. All message which is received is rebroadcasted even they do not participate in route setting. It brings bandwidth waste and traffic overhead. In this paper, we propose a belt-zone concept and a selection sequence method for decreasing the number of unnecessary control messages and increasing network efficiency.

2 AODV Routing

AODV routing protocol supports the multi-hop routing among mobile nodes for establishing and maintaining an ad hoc network. In AODV, a node requests a route only when it is needed, and the other case of nodes do not need to maintain routing table. To send a message from the source to the destination, the source node initiates a route discovery procedure. A RREQ message is flooded through the network until it reaches to the destination or it reaches to a node that knows the route to the destination.

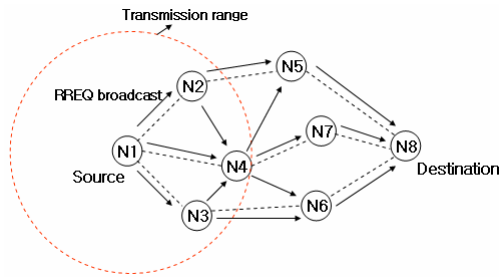


Fig. 1. AODV Routing Protocol Route Request

Fig. 1. shows RREQ messages flooding. Nodes received. RREQ messages check if node itself is a destination node. If the node is not the destination node, increases RREQ messages hop count by 1 and broadcasts RREQ messages to neighbor nodes. While transmitting RREQ messages, intermediate nodes save RREQ packets broadcasted node's address in routing table and use in reverse path. A duplicated RREQ message is dropped.

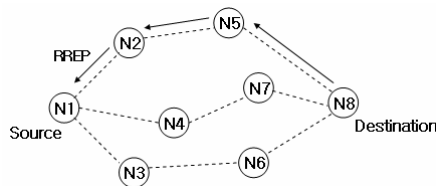


Fig. 2. AODV Routing Protocol Route Reply

A route reply (RREP) control message from the destination is unicast back to the source along the reverse path. Fig. 2 shows a transmission process of a RREP message from destination to source node. When the RREP is routed back along the reverse path, all nodes on this route set up a forward path by pointing the node that transmit the RREP. These forward route entries indicate the route to the destination node. Through this procedure, the route is made. In AODV, by flooding the RREQ message, a lot of RREQ messages have to be transmitted. This derives large overhead and fast power consumption.

3 A Belt-Zone Selection Method

The purpose of belt-zone method is to decrease the number of unnecessary RREQ messages. All nodes that receive the RREQ control message need not rebroadcast the message, but only the nodes within a selected zone rebroadcast it and set the routing path. When each node receives the RREQ, the node checks the signal strength if it is within in-bound and out-bound signal strength threshold. When the state is true, then it is in selected zone and rebroadcasts the message. This can decrease the number of the RREQ control messages. The node that rebroadcasts the RREQ message and establishes a stable routing path is considered a member node of selected belt-zone.

3.1 The Concept of Belt-Zone

A belt-zone(BZ) is a scope of area. The nodes in the selected area rebroadcast the RREQ message received from a sender. It is included in belt-zones which is a logical concentric area within transmission range from the sender. In fig. 3, between $SSTR_{in}$ (in-bound Signal Strength Transmission range) and $SSTR_{out}$ (out-bound Signal Strength Transmission range) is set as a selected BZ. Nodes inside the BZ rebroadcast a RREQ message that is received from A(Sender) and nodes outside of the BZ abandon it. In fig. 3, node A broadcasts a RREQ, then node B in the selected belt-zone rebroadcast the RREQ and nodes C, D, E, F abandon it. The signal strength of node B detects is weaker than $SSTR_{in}$ and stronger than $SSTR_{out}$, therefore node B is included in the belt-zone and rebroadcasts RREQ message to neighbor nodes as a new sender. Detected signal strength by node C is stronger than $SSTR_{in}$ and it can decide itself as out of belt-zone, therefore node C abandons the message. BZ can be decided as follow (1).

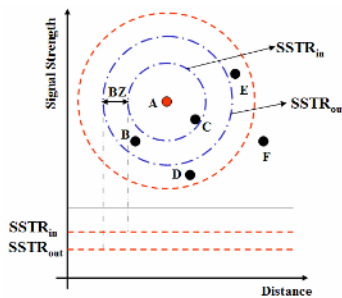


Fig. 3. The measurement of signal strength

$$SSTR_{in} \geq \text{received SS} \geq SSTR_{out} . \quad (1)$$

In fig. 4, when the source node A sends RREQ messages to find the destination node N. Node B, C, D, E, F are in the transmission area of node A depend on measured signal strength. If a node is aware that it is between $SSTR_{in}$ and $SSTR_{out}$, the rebroadcasts the RREQ message. In fig. 4 node B and E are decided and rebroadcast the message, but node D, C, F abandon the message. The rebroadcasted RREQ message sent from node E is received by nodes C, D, F, I, J, and K. Only node J and K belongs to the BZ for the message, therefore these nodes rebroadcast it.

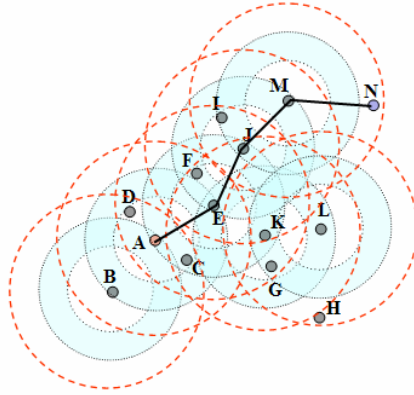


Fig. 4. A route discovery procedure using the belt-zones

By the similar sequence, node J, M relay the message to destination node N.

3.2 Belt-Zone Extension

Nodes broadcast RREQ control messages only when the node belongs to the selected BZ. If a route discovery procedure is failed, the source node extends BZ scope and restart a new route discovery procedure to find the destination. Function (2) is the first belt-zone extension case.

$$SSTR_{out(N)} = SSTR_{out(P)} - BZ_{width} . \quad (2)$$

Where $SSTR_{out(P)}$ is previous signal strength, $SSTR_{out(N)}$ is the new signal strength threshold, and BZ_{width} is extension parameter. In the first expansion, the area is extended to the direction of the outside.

After first time belt-zone expansion, the route discovery procedure is repeated. If finding the destination is again fail, then BZ scope area is re-extended. Function (3) shows the case of second extension of BZ area. In this case, the broadcast area is extended to in-bound direction.

$$SSTR_{in(N)} = SSTR_{in(P)} + BZ_{width} . \quad (3)$$

As the expansion result, the BZ scope includes whole transmission area from the sender as shown in fig. 3.

4 Performance Evaluation

4.1 The Belt-Zone Model

To select efficient belt-zone, we use two models. One is divided as three zones and the other is divided as four zones.

In first model, we simply divide whole transmission area as three logical concentric BZ areas. BZ_1 is the area in which the signal strength is from $SSTR_0$ to $SSTR_1$. BZ_2 is the area in which the signal strength is from $SSTR_1$ to $SSTR_2$. BZ_3 is the area in which the signal strength is from $SSTR_2$ to $SSTR_3$.

In second model, we select the big inner area from the sender as BZ_0 and remained outside transmission area will be divided for three BZ areas. BZ_0 is the area in which the signal strength is below $SSTR_0$. BZ_1 is the area in which the signal strength is from $SSTR_0$ to $SSTR_1$. BZ_2 is the area in which the signal strength is from $SSTR_1$ to $SSTR_2$. BZ_3 is the area in which the signal strength is from $SSTR_2$ to $SSTR_3$.

By simulation results, we find that BZ_1 of fig. 5 and BZ_0 of fig. 6 are seldom selected as selected BZ. Therefore for effective analysis, we do not select those two area as selected BZ in the future experiments.

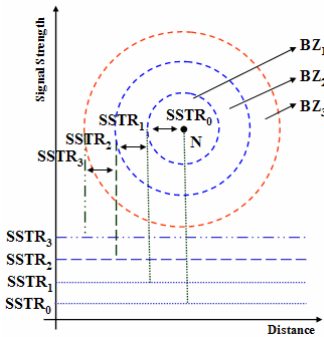


Fig. 5. First belt-zone model for simulation

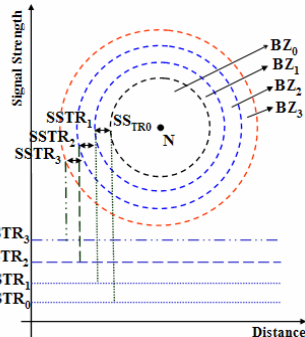


Fig. 6. Second belt-zone model for simulation

4.2 Simulation Environment

Simulation environment consists of the number of nodes = 5, 13, 25, 50 per unit area, and use the area of 2000 x 2000. Each node is distributed randomly. For each parameter, 100 simulations are done. To find out the most effective first selection belt-zone area, by the number of RREQ messages, the success rate of transmission is monitored depending on the changing of node density.

Simulation result of first model shows that only BZ_3 has possibility to be selected as belt-zone as shown in fig. 7-8. We can realize that BZ_3 has 100% success rate of transmission and the number of RREQ messages are far fewer than AODV as shown in fig. 9-10.

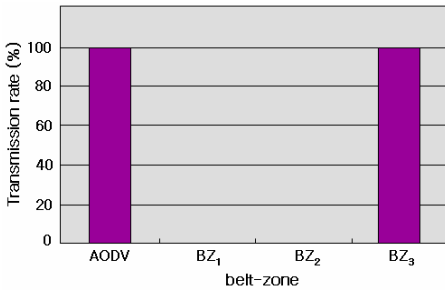


Fig. 7. The Success rate of transmission in first model when $nd=5$ (low density)

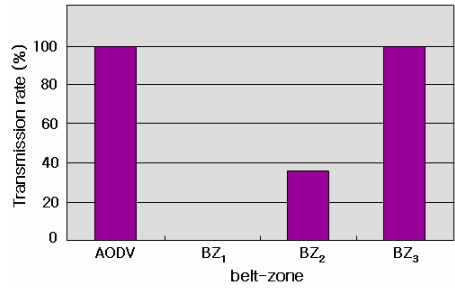


Fig. 8. The Success rate of transmission in first model when $nd=25$ (high density)

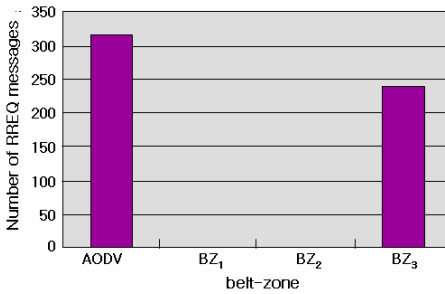


Fig. 9. The Number of RREQ message in first model when $nd=5$ (low density)

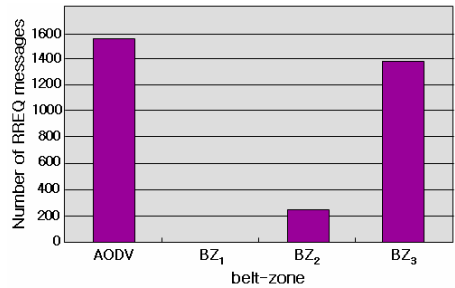


Fig. 10. The Number of RREQ message in first model when $nd=25$ (high density)

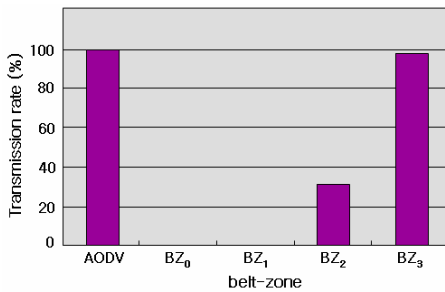


Fig. 11. The Success rate of transmission in second model when $nd=5$ (low density)

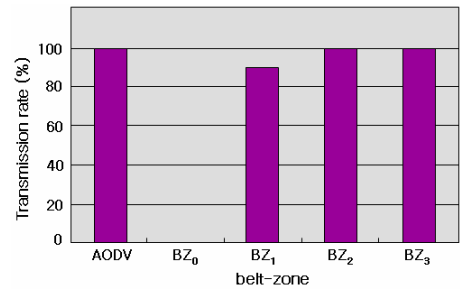


Fig. 12. The Success rate of transmission in second model when $nd=25$ (high density)

The simulation result of [Fig. 11-12] shows that the success rate of transmission in low node density and high node density case. Regardless of node density, success transmission rate of BZ₀ is 0%.

Fig. 13-14 Show the number of RREQ control messages with traditional AODV and proposed belt-zone method in low density case and high density case. As we can

see BZ_3 is found as the most efficient area for the condition of low density. For high-density condition, BZ_2 is found as the most efficient area. We can say that first selected BZ must be different depend on the node density of BZ.

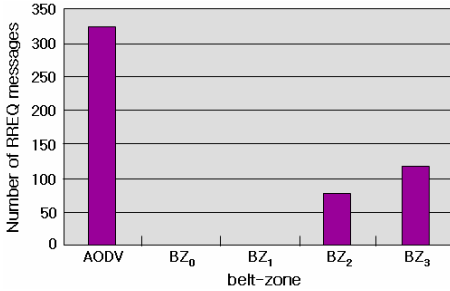


Fig. 13. The Number of RREQ message in second model when $nd=5$ (low density)

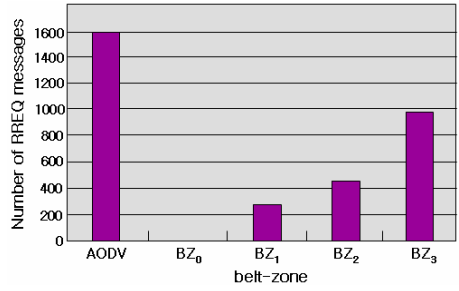


Fig. 14. The Number of RREQ message in second model when $nd=25$ (high density)

Density decision function of the initial belt-zone selection is as follows:

$$\begin{aligned}
 BZ &= f(N), \\
 &\text{if } (N \leq 10) \text{ then } BZ_3, \\
 &\text{if } (N > 10) \text{ then } BZ_2,
 \end{aligned}
 \tag{4}$$

where N is the number of nodes in a unit area.

According to above density decision function, first BZ selection will be set up and RREQ message will be broadcasted. As expansion of belt-zone selection method, first the most suitable belt-zone must be selected. If route-establishment fails, then the belt-zone area must be extended.

In low density case, BZ_2 is added as extension to BZ_3 . Fig. 15 shows despite belt-zone is extended, total number of RREQ message is fewer than that of AODV. Fig. 16 shows transmission success rate of both are 100%.

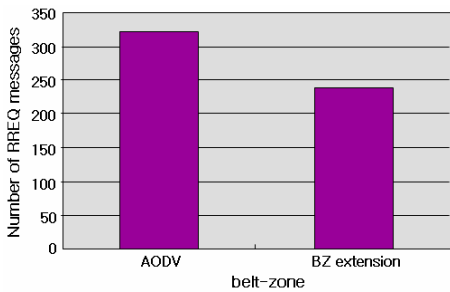


Fig. 15. The number of RREQ message of AODV and that of extended BZ in low density



Fig. 16. Transmission success rate of AODV and that of extended BZ in low density

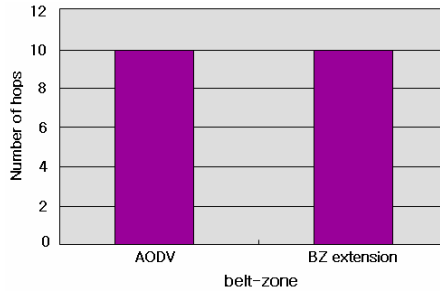


Fig. 17. The number of AODV hop and that of extended BZ hop in low density

Fig. 17 shows that hop counts are the same even though the methods are different.

As the results of simulation, we confirm that the proposed BZ method can give apparent improvement of performance.

5 Conclusions

We proposed a belt-zone model to select the most suitable subset of communication area for decreasing RREQ messages. The proposed BZ model divides the whole transmission range from the send as subsets of concentric areas. The key idea is that a belt-zone is selected as a rebroadcast area, for control message then only the nodes in the zone attend to establish route. If the route searching process fails, then the belt-zone area is extended. The primary selection of a belt-zone depends on node density. In the case of low-density, selecting out-bound area is prior. Extension of the area is done to in-bound areas. In the case of high-density, selecting intermediate area is prior. Extension of the belt-zone is done first to out-bound, then next to in-bound.

By applying our proposed belt-zone method to AODV networks, we can observe that a lot of control messages are decreased on various load situations.

References

1. E. M. Royer and C. -K Toh, "A Review of Current Routing Protocols for Ad-Hoc Mobile Wireless Networks." IEEE Personal Communications. pp. 46-55, April 1999.
2. Charles E. Perkins and Bhagwat, "Highly dynamic destination sequenced distance-vector routing(DSDV)for Mobile Computers," ACM SIGCOMM, Oct., 1994.
3. T. Clausen, P. Jacquet, A. Laouiti, P. Minet, P. Muhlethaler, A. Qayyum and Laurent Viennot, "Optimized Link State Routing Protocol", Internet Draft, IETF MANET Working Group, draft-ietf-manet-olsr-07.txt, December 2002.
4. J. Li, H. Kameda, and Y. Pan, "Study on Dynamic Source Routing Protocols for Mobile Ad-Hoc Networks," Proceedings of Workshop on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks 2003 (WiOpt' 03), pp. 337-338, INRIA, Sophia-Antipolis, France, March 3-5, 2003.
5. V. Park and M. Corson, "Temporally Ordered Routing Algorithm (TORA) Version 1 Functional Specification", Internet Draft, IETF MANET Working Griop, draft-ietf-manet-tora-spec-02.txt, October 1999.

6. S. Roy and J.J. Garcia-Luna-Aceves, "An Efficient Path Selection Algorithm for On-Demand Link-State Hop-by-Hop Routing", Proc. IEEE International Conference on Computer Communications and Networks (ICCCN), Miami, FLorida, October 14-16, 2002.
7. J.J. Garcia-Luna-Aceves, M. Mosko, and C. Perkins, " A New Approach to On-Demand Loop-Free Routing in Ad Hoc Networks, " Proc. Twenty-Second ACM Symposium on Principles of Distributed Computing (PODC 2003), Boston, Massachusetts, July 13--16, 2003.
8. Z. J. Hass and M. R. Perlman, "The Zone Routing Protocol (ZRP) for Ad Hoc Networks", Internet Draft, IEFE MANET Working Group, draft-ietf-manet-zone-03.txt, March 2000.
9. Widmer, J., Mauve, M., Hartenstein, H., FuBler, H. "position-Based Routing in Ad-Hoc Wireless Networks." In: The Handbook of ad Hoc Wireless Networks. Hrsg., Ilyas, Mohammad. Band, Auflage Boca Raton, FL. CRC Press, 2002, S. 12-1 – 12-14.
10. C. K. Toh, "Long-lived ad hoc routing based on the concept of associativity," Internet draft, IETF, Mar., 1999.
11. Mary Wu, Younrag kim, chonggun kim, "A Path Fault Avoided RPAODV Routing in Ad Hoc Networks", KIPS 11-C(7) 2004.

A VLSM Address Management Method for Variable IP Subnetting

SeongKwon Cheon¹, DongXue Jin², and ChongGun Kim^{2,*}

¹ Division of Computer Information, Catholic Sangji College,
Andong, Gyeongbuk, Korea
skcheon@csangji.ac.kr

² Dept. of Computer Engineering, YeungNam University,
Gyeongsan, Gyeongbuk, Korea
donghak@yumail.ac.kr, cgkim@yu.ac.kr

Abstract. IPv6 have been examining at the next IP address standard. But IPv4 have to be used for a while by the following reasons: tremendous cost and efforts for converting to IPv6. One of the serious problems of the IPv4 addressing structure is the fact that is a shortage of IP addresses. The address shortage is derived by lots of unused addresses during IP distribution and IP subnetting design. We propose an effective subnet IP address calculation method on VLSM. Also, with the proposed subnet IP address management method, a web based subnet address management system is introduced. The web-based subnet IP management system offers convenience in VLSM-based subnetting. The proposed VLSM calculation method can give a simple and effective IP management.

1 Introduction

The Internet is growing at an incredible rate. Every system that is connected to the Internet requires an IP address, which can uniquely be identified as individual system. The IP address on a computer must be unique on a world-wide basis and duplicates are not allowed. By the shortage of IPv4 addresses, IPv6 have been derived and is examining at the next IP address standard [1]. But, to convert IPv4 address to IPv6 for all world wide Internet nodes need tremendous cost and efforts [2, 3]. That is why IPv4 is currently used at the standard IP address of Internet and will be used for a while.

The IPv4 address system showed only a few problems during the early years of the Internet, but its weaknesses began to emerge as the Internet grew at a very fast rate. One of prominent problems of the current IP address system is the fact that many addresses are wasted during allocation of IP addresses. Some research results in this field are Virtual IP, CIDR, VLSM, etc.[12, 13, 14, 15], and a long term solution to overcome problems is the next-generation Internet address system, IPv6, which was proposed by the IETF [4].

* Correspondence author.

In 1985, the Subnet concept had been announced from the ‘RFC 950, Internet Standard Subnetting Procedure’ in order to solve the problems of IP address shortage [5]. The advantage of subnetting is that network traffic load can be reduced and structuring the internal network into multi-level hierarchy can increase security. The Subnet is realized through the Subnet Mask. In general, a network is divided into equal-sized subnets using a single subnet mask. When a single subnet mask is used, the number of hosts that can be attached for each subnet becomes equal. In this case, the difference between the number of IP addresses allocated and the number of actually used ones on the subnets becomes the number of wasted ones.

As another solution, VLSM(Variable Length Subnet Mask) address allocation was proposed in 1987 at the IETF with RFC 1009 [6]. Using multiple subnet masks, it reduces the waste of address space by generating different-sized subnets in proportion to the proper number of connected hosts to the subnet. However, because VLSM uses multiple subnet masks, the subnetting process is very complicated and management of subnet IP address is difficult. Many managers tend to avoid the use of VLSM due to its complicate management and enormous efforts.

In this paper, a straight forward VLSM calculation method and efficient VLSM subnet IP address management method is proposed. A prototype web-based subnet management system is designed and implemented.

2 IP Address Allocation

In order to exchange messages over the Internet, each host must be identified by a unique IP address. Inside at a LAN, allocated IP addresses by NIC must be managed by subnetting.

2.1 IPv4 Address Systems and Address Class

Currently, the Internet is mainly using the IPv4 address system and an IP address can be divided into the network identifier(netid) and the host identifier (hostid). The network identifier represents the network to which a particular computer belongs to and the host identifier represents each host or router within the network.

IP address is classified by 5 classes, A, B and C classes as the network identifier and D, E classes for special purposes. The class-based two-level address architecture has faced many problems with the rapid growth of the Internet. Class B address is almost completely exhausted. There are not many organizations which can efficiently use the 16 million class A addresses. The C class network, which can support 256 IP addresses, is too small.

CIDR was proposed in the early 1990's and developed in September of 1993, CIDR was distributed through RFC 1517, 1518, 1519 and 1520. CIDR cooperates with bit masks. The number of bits that designates the network and the number of bits that designates the host may be different according to the length of the address prefix. The length of the prefix can be determined by the address class or using CIDR(Classless Inter-Domain Routing). RFC 1878 lists the 32 possible prefix values [7].

2.2 Subnet Mask and Subnet

The IP address is divided into a network part and a host part. However, it is inefficient to apply only this two-level concept to the complex and multifarious network structures. For example, an institution that has a B class address that has approximately 65,000 IP addresses, if the subnet concept is not applied, all hosts will be on the same level and thus the traffic overhead will be enormous.

Subnets which can decrease the network load and increase security by hierarchically structuring the internal network can be generated by the subnet mask. The advantage of subnetting is that the network can be derived into a proper manageable size subnets [8]. Figure 1 shows an example of subnet by applying a 255.255.255.0 subnet mask to the IP address.

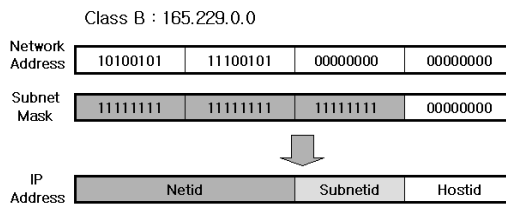


Fig. 1. Extracting a subnetid by subnet mask

In figure 1, the subnetid part has 8 bits, which means the network is divided into 254 the same sized subnets. Each subnet can connect up to 254 hosts. If we want more subnets, then the number of hosts per subnet decreases. On the other hand, if the number of subnets decreases then the number of hosts per subnet increases. The structure of subnets cannot give any effect to outside and the whole local network is recognized as a single network by the network identifier.

2.3 VLSM(Variable Length Subnet Mask)

VLSM(Variable Length Subnet Mask) is using multiple subnet masks on a single network. With this technique, different-sized subnets can be generated in one local network. Therefore, each subnet can have proper number of hosts.

Figure 2 shows an example of different-sized addresses within a class B network using VLSM. Since subnets ①, ② and ③ are Point-to-Point connections, only two IP addresses are required. Applying subnet mask 255.255.255.252, a small subnet with 4 host ID's are generated and allocated. The 4 host ID's include the subnet and broadcast addresses. Subnet mask 255.255.255.128 is applied to subnet ④ and a subnet with 128 hosts is allocated. Subnet mask 255.255.255.192 is applied to subnet ⑤ and a subnet with 64 hosts is allocated. Finally, Subnet mask 255.255.255.0 is applied to subnet ⑥ and a subnet with 256 hosts is allocated.

Table 1 shows the subnetting of 165.229.0.0 class B network by applying three subnet masks. Four subnets 165.229.0.0 / 165.229.64.0 / 165.229.128.0 / 165.229.192.0 are generated when 165.229.0.0 is subnetted using subnet mask 255.255.192.0. Also, eight subnets 165.229.0.0 / 165.229.32.0 / 165.229.64.0 / 165.229.96.0 / 165.229.128.0

/ 165.229.160.0 / 165.229.192.0 / 165.229.224.0 are generated when subnet mask 255.255.224.0 is used. Likewise, 32 subnets including 165.229.0.0 / 165.229.8.0 are generated when subnet mask 255.255.240.0 is used. However, special attentions are needed in order to avoid confliction among IP addresses, if 165.229.0.0 is used for generating subnets by applying subnet mask 255.255.192.0, subnets 165.229.0.0 and 165.229.32.0 generated by subnet mask 255.255.224.0 cannot be used. Also, 8 subnets generated by subnet mask 255.255.248.0, which are subnets 165.229.0.0 through 165.229.56.0 cannot be used as well.

The probability of problems occurring is greater as the number of subnet masks used in a network increases. In general, 2 or 3 subnet masks are recommended on VLSM.

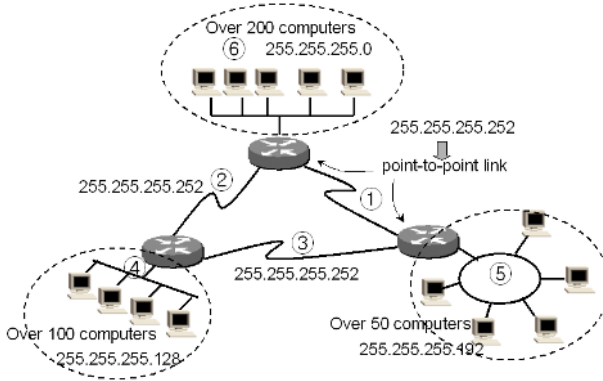


Fig. 2. Different-sized subnets

Table 1. Subnetting with three subnet mask

First subnet mask: 255.255.192.0 CIDR/18
Second subnet mask: 255.255.224.0 CIDR/19
Third subnet mask: 255.255.248.0 CIDR/21

Subnet Mask	Starting Host	Last Host	Broadcast
165.229.0.0	165.229.0.1	165.229.63.254	165.229.63.255
165.229.0.0	165.229.0.1	165.229.31.254	165.229.31.255
165.229.0.0	165.229.8.1	165.229.7.254	165.229.7.255
165.229.8.0	165.229.16.1	165.229.15.254	165.229.15.255
165.229.16.0	165.229.24.1	165.229.23.254	165.229.23.255
165.229.24.0	165.229.32.1	165.229.31.254	165.229.31.255
165.229.32.0	165.229.32.1	165.229.63.254	165.229.63.255
165.229.32.0	165.229.32.1	165.229.39.254	165.229.39.255
165.229.40.0	165.229.40.1	165.229.47.254	165.229.47.255
165.229.48.0	165.229.48.1	165.229.56.254	165.229.56.255
165.229.56.0	165.229.56.1	165.229.63.254	165.229.63.255
165.229.64.0	165.229.64.1	165.229.127.254	165.229.127.255
165.229.64.0	165.229.64.1	165.229.63.254	165.229.63.255
...

3 A Variable-Length Subnet IP Address Management Method

To make easy subnet addresses on VLSM, we need to know simply which part of IP addresses is already used and which ones are available.

3.1 A Concept of IP Subnet Address Allocation on VLSM

The sizes of the subnets within a network are different from each other according to the subnet masks applied. The number of hosts that can be allocated for each subnet decreases as the number of subnets increases, and as the number of subnets decreases, the number of hosts that can be allocated for each subnet increases.

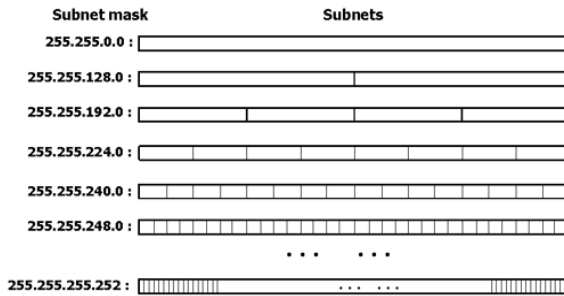


Fig. 3. Subnet mask and subnets

Figure 3 briefly shows the subnet size produced by each subnet mask in a class B network permitted from the NIC. In the class B, the basic subnet mask is 255.255.0.0 as in the figure and uses the entire network as a single flat subnet. If the subnet mask 255.255.128.0 is applied, it can be allocated into two subnets and if subnet mask 255.255.192.0 is applied, it can be allocated into 4 subnets. Likewise, if the subnet mask is applied as 255.255.255.252, a subnet with 4 host addresses can be allocated into 16,384 subnets.

The basic concept of efficient VLSM-based IP address management which is introduced in this paper is that whenever a subnet is allocated, the domain of that particular subnet is marked 'used,' and the same domain using different subnet mask is set as 'unusable.' It prevents conflict among the subnet addresses allocated by different size subnet masks [9].

If a subnet using subnet mask 255.255.224.0 is allocated, the first subnets using subnet masks 255.255.192.0, 255.255.128.0 and 255.255.0.0 are marked 'unusable.' Also, the first two subnets using subnet mask 255.255.240.0 and the first four subnets using subnet mask 255.255.248.0 cannot be used. Likewise, the first 1,024 subnets using subnet mask 255.255.255.252 should also be marked 'used.'

Figure 5 shows the result of allocating a subnet with subnet mask 255.255.192.0, and figure 6 shows the result of allocating five more subnets with subnet mask 255.255.240.0. In this manner, when some IP addresses are needed to allocate, we can allocate by starting from the left in the usable subnet domain. Also, because the

allocated IP addresses are concentrated more on the left, the remained usable IP addresses are used to generate subnets that can support large sized subnets.

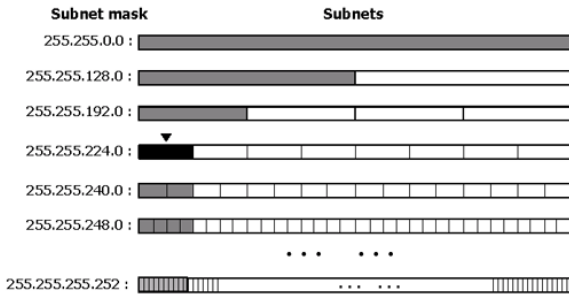


Fig. 4. First stage of Subnetting

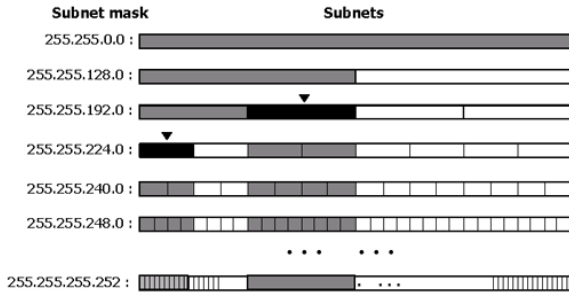


Fig. 5. Second stage of subnetting

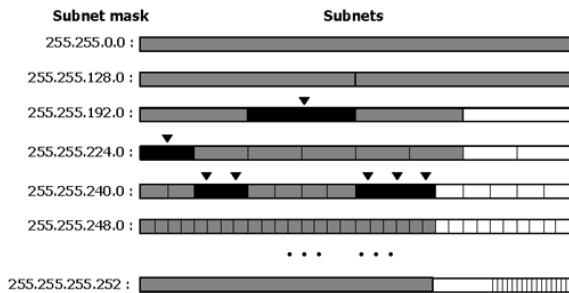


Fig. 6. Third stage of subnetting

3.2 An Array System for the Subnet Management

The minimum number of hosts of a subnet which includes networked and a broadcast address should be four[11]. The proposed management structure uses 4 IP addresses

as the minimum unit, and is applied using subnet mask 255.255.255.252. The ‘0’ shows ‘usable’ and a CIDR prefix number which represents subnet mask shows ‘already used’ or ‘unusable’.

Figure 7 shows the initial IP address management array for a class B network address. Because four IP addresses are managed by one element, the management array for a class B network must have $65,536 / 4 = 16,384$ elements. The array is initialized with 0’s that means ‘usable.’

Every time when a subnet allocation is needed, the elements of the management array must be checked and array blocks marked ‘usable’ can be allocated. The value of the elements must be CIDR prefixes values of the subnet mask. This value helps that the system needs not to check all the elements in the array to find ‘usable’ blocks. We define domain as group of array elements. The CIDR value gives information of domain size already used.

Let us assume that p is the CIDR prefix of a subnet mask to be used. Then, the number of IP addresses per subnet is 2^{32-p} . However, since each element of the management array manages four IP addresses simultaneously, 2^{30-p} elements manage 2^{32-p} IP addresses. Therefore, the interval between the domain that which is group of elements that manage the subnet addresses is 2^{30-p} and in order to check the usage, the 0th, 2^{30-p} th, $2 * 2^{30-p}$ th, $3 * 2^{30-p}$ th ... elements needs to be checked by order.

As in figure 4, figure 8 shows the process of finding an available element range of the array and setting it by the corresponding value. The gray elements are the ones to be checked for using later on when p is.

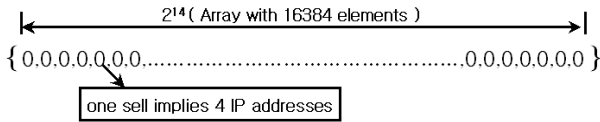


Fig. 7. Initialization of subnet management array

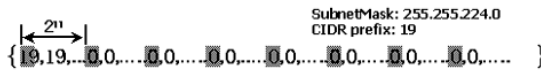


Fig. 8. Management array of first stage of subnetting

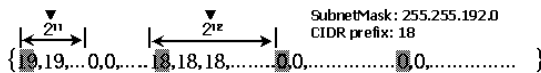


Fig. 9. Management array of second stage of subnetting



Fig. 10. Management array of third stage of subnetting

As in figure 5, figure 9 shows the result of allocating a subnet with value 18. Also, as in figure 6, figure 10 shows five more subnets allocated with value 20.

3.3 Information Retrieved from the Management Array

The array information has been saved as the subnet management information. Whereas, by analyzing the element values of the management array, we can get exact IP subnet addresses.

One element of the management array represents four IP addresses. Let's Assume that the subnet CIDR prefix value is p , and the value is included in the i th position in the management array. The following subnet information can be easily calculated using values p and i .

- Subnet mask
- Subnet address
- Start and end addresses of the host
- Broadcast address

Here, the subnet mask can be easily determined from the CIDR prefix value p .

For the subnet address, in the case of a class B subnet address, the first two bytes are decided and only the last two bytes need to be determined. The third and fourth byte of class B IP addresses are incremented by 1 starting from 0.0. Whenever the fourth byte is incremented beyond 255, the third byte is incremented by one and the fourth byte is again reset to 0. However, because every element of the management array represents four IP addresses, if the i th element contains a subnet size, the IP address of the corresponding subnet becomes $4i$ th IP address. We can show it as mathematical expressions.

$$[(4 * i) / 256].[(4 * i) \% 256]. \text{-----} \tag{1}$$

The broadcast address is equal to subtracting 1 from the subnet address of the next position. Since the subnet addresses exists in 2^{30-p} intervals, the third and fourth bytes can be expressed as the following,

$$[(4 * (i + 2^{30-p}) - 1) / 256].[(4 * (i + 2^{30-p}) - 1) \% 256]. \text{-----} \tag{2}$$

Also, the available beginning host address of the subnet is equal to adding 1 to the subnet address, that is,

$$[(4 * i + 1) / 256].[(4 * i + 1) \% 256]. \text{-----} \tag{3}$$

The available last host address of the subnet is equal to subtracting 1 from the broadcast address, that is,

$$[(4 * (i + 2^{30-p}) - 2) / 256].[(4 * (i + 2^{30-p}) - 2) \% 256]. \text{-----} \tag{4}$$

The element values are CIDR prefix values of the used subnet mask. The beginning and end of one subnet can be identified. The subnet information that can be determined with equations (1), (2), (3) and (4). For example, in figure 11, if each value of each element in the management array is read sequentially, the value of the 0th ele-

Subnetting with VLSM

[\[Log out\]](#) [\[Initial\]](#)

Welcome! Member **John Smith**

Your network address is Class **B**.

:: Network address : 165 . 229 . 0 . 0

:: Information of sebnets:

New subnets

No.	SubnetMask	SubnetAddress	Start Host	End Host	Broad Cast	Hosts/Subnet
-----	------------	---------------	------------	----------	------------	--------------

All subnets

No.	SubnetMask	SubnetAddress	Start Host	End Host	Broad Cast	Hosts/Subnet
-----	------------	---------------	------------	----------	------------	--------------

:: New create ::
by hosts/subnet

Select Hosts/Subnet: Number of subnet:

:: New create ::
by subnetmask

Select SubnetMask: Number of subnet:

Fig. 12. Page of subnetting

The subnet allocation information is divided into two sections. The 'new generated subnet' section displays the information of the newly generated subnets when a new subnet is allocated, and the 'all subnets' section displays the information of all the subnets allocated before. Each subnet information includes 5 fields that which are the subnet mask, subnet address, start host, end host and broadcast, and the number of hosts per subnet is calculated and displayed.

5 Conclusions

In this study, an easy and efficient management method for variable IP subnet addresses based on VLSM is proposed and a prototype web-based subnet management system is also designed and implemented.

The proposed calculation and management method for subnetting IP address is used for IP management system for VLSM-based LAN. The designed and implemented web-based subnet management system is straight and efficient for network IP designer. This system would be a solution for easy management for VLSM subnetting environment. The derived results may help for decreasing IPv4 waste during IP address design and maintenance until IPv6 becomes the main IP address on whole Internet.

References

1. R. Hinden, S. Deering, "Internet Protocol Version 6 (IPv6) Addressing Architecture", RFC 3513, Nokia, Cisco Systems, April 2003
2. J. Bound, Ed., "IPv6 Enterprise Network Scenarios", Hewlett Packard, June 2005
3. J. Hagino, K. Yamamoto, "An IPv6-to-IPv4 Transport Relay Translator", IJ Research Laboratory, June 2001
4. S. Bradner, A. Mankin, "The Recommendation for the IP Next Generation Protocol", RFC 1752, Harvard University, ISI, January 1995
5. J. Mogul, J. Postel, "Internet Standard Subnetting Procedure", RFC 950, Stanford, ISI, August 1985
6. R. Braden, J. Postel, "Requirements for Internet Gateways", RFC 1009, ISI, June 1987
7. T. Pummill, B. Manning, "Variable Length Subnet Table For IPv4", RFC 1878, Alantec, ISI, December 1995
8. F. Baker, Editor, "Requirements for IP Version 4 Routers", RFC 1812, Cisco Systems, June 1995
9. Dong Hak Kim, Seong Kwon Cheon, Mary Wu, Chong Gun Kim, "Effective subnet IP address allocation by using VLSM", The Korea Multimedia Society, Conference at Autumnfield, Vol. 5, No.3, PP. 65-68, 2002.11
10. Y. Rekhter, T. Li, Editors, "An Architecture for IP Address Allocation with CIDR", RFC 1518, T.J. Watson Research Center, IBM Corp, cisco Systems, September 1993
11. SeongKwon Cheon, DongXue Jin, Mary Wu, ChongGun Kim, "An Effective VLSM subnet IP address allocation and management method", THE 13 JOINT CONFERENCE ON COMMUNICATIONS AND INFORMATION (JCCI 2003), Poster Session III Network, P-III-14.1~4
12. Y. Rekhter, B. Moskowitz, D. Karrenberg, G. J. de Groot, E. Lear, "Address Allocation for Private Internets", RFC 1918, Cisco Systems, Chrysler Corp, RIPE NCC, Silicon Graphics, Inc., February 1996
13. Classless Inter-Domain Routing (CIDR), "http://user.chollian.net/~son6971/Internet_address/ch5/ch5.htm"
14. Robert Wright, "IP Routing Primer", Cisco Press, 1998
15. J.D. Wegner, Robert Rockell, "IP Addressing and Subnetting", Syngress, 2002

SDSEM: Software Development Success Evolution Model

Haeng-Kon Kim¹ and Sang-Yong Byun²

¹ Department of Computer Information & Communication Engineering,
Catholic University of Daegu, Korea
hangkon@cu.ac.kr

² Division of Computer Information & Communication Engineering,
CheJu National University, Korea
byunsy@cheju.ac.kr

Abstract. Each company of organization should develop its own model or tailor the above models to make them suitable to its unique environment such as product or technology domain, scale of business or organization and cultural environment, etc for the practical application. In this paper, we introduces a case in which organizational and technical capability was reinforced based on our own process capability improvement model which is named SDSEM (S/W Competence Reinforcement Model) to improve S/W development strength in a corporate, which manufactures varieties of consumer electronics products which are embedding controller S/W as its brain and in which large-scale development organization has multi-site development environments.

We evaluated SDSEM as a very practical but limited model against our goal by introducing and applying to business units.

1 Introduction

The importance of software in consumer electronic products is enormously increasing as time goes by. new IT technology is being applied to electronic products and users also want to have varieties of functions from electronic products beyond their old and traditional functions. We are forced to expand S/W business area from a simple embedded S/W driving the traditional single component to network S/W which expands connectivity among products, application S/W which presents differentiated-advanced functions, and even to S/W services which provide integrated solution and service. Especially, it's unquestionable the importance of S/W is more stressed out under the environment of Digital Convergence which is the future keyword being claimed[1,2].

We also started to define and manage S/W development process under product development process and its lifecycle to systematically develop embedded S/W that is becoming more complex and monstrous. By introducing S/W process for the past several years, we researched and accumulated technologies about experiences and methods for numerous process modeling and management necessary in electronics business.

This study figures out important viewpoints based on actual application cases for the insight of product development process from the viewpoint of embedded S/W business domain. This study also reviews process modeling and management method

necessary for efficient product development. It represents the research direction for developing practical process model based on some actual examples.

This study can be a reference as industry experiences improving the capability of organization by tailoring process maturity model in an enterprise, and it will be a contribution as an industry feedback to a process model study.

2 Backgrounds

2.1 Process Maturity Model

The Capability Maturity Model Integration (CMMI) project is a collaborative effort to provide models for achieving product and process improvement. The primary focus of the project is to build tools to support improvement of processes used to develop and sustain systems and products. The output of the CMMI project is a suite of products, which provides an integrated approach across the enterprise for improving processes, while reducing the redundancy, complexity and cost resulting from the use of separate and multiple capability maturity models[3]. IS 15504 is a suite of standards for software process assessment currently an international standard. This International Standard provides a framework for the assessment of software processes. It can be used by organizations involved in planning, managing, monitoring, controlling, and improving the acquisition, supply, development, operation, evolution and support of software[4].

2.2 Needs for Process Improvement

It becomes a necessary field for IT business to improve management capability of IT resources to invest them more efficiently and maximize their effect.

IT has emerged as a core field to reengineer and develop business process and to improve business process through utilization of computer. Gigantic companies such as IBM, Ford, and GE obtain excellent result of more than 80% from reengineering field rather than business improvement effects by computer [4].

2.3 SDSEM: Goal and Scope

The business needs of CE segment are rapidly changing, especially with Software fields. We can roughly classify it as 3 categories.

The first one is the change of market environment. The connectivity expansion among product groups are increasing, and the barriers among product groups are collapsing due to technical unification and convergence among product groups, and a specific technology necessary to a specific product is expanding its utilization scope horizontally across entire product groups. Software technologies are in the center of these digital convergences.

The second one is the need for continuous business planning updates to cope with tough business conditions. In accordance with the business plan of pursuing world best, the premium features of product groups are very important issues. The functions of these premium products should implement new features and new functions which can't be found with the traditional products and they should also guarantee high performance. These functions usually have very close relationship with software.

The third one is that the value of software in a product itself is increasing. We can know that indirectly by hearing the claims caused by software problems which are dramatically increasing nowadays.

Eventually, Software with differentiated quality and functions leads to the competitiveness of a product. It means Software development competence becomes the competitiveness of a company. For this reason, reinforcement method of S/W development competence in an enterprise is a very important issue, and it's necessary to develop a practical methodology to resolve this issue. The 115504 and CMMI present framework for this, but it can't provide a practical method. That's why the development of methodology holistically taking into account internal and external environments and cultural factors an enterprise faces with is indispensable. Figure 1 shows the goal and scope of the SDSEM. Models help us focus, visualize, and reason by abstracting away parts of the overall software situation and by making simplifying assumptions. SDSEM which is suggested in this paper is developed on the concept of the same reason. We need visualized thing which can be understood and execute it easily. Dr. Boehm suggested process models for visualizing and reasoning about software project should be and is going, product models for that the product should be doing, property.

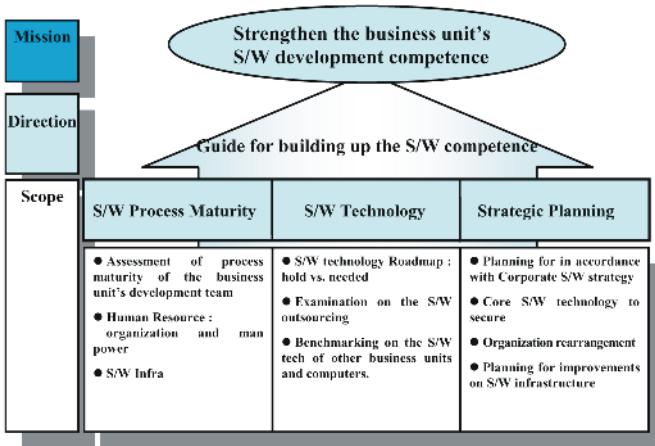


Fig. 1. Goal and scope of the SDSEM

3 Model Design

3.1 Overall Structure of SDSEM

We defined the goal of model and detail procedures to make our model to be practical and effective about innovation of our business units for the reinforcement of S/W development competence. We need to define the top level, abstract model as a common factor which can be utilized for the improvement of software development competence in each of business units. The goal of SDSEM is to present mid to long-term

software innovation direction by analyzing current situation of the business unit to empower technical and organizational strength.

SDSEM is composed of 5 sub models as shown in Figure 2 and it includes Domain Model, Process Evaluation Model, SOP(S/W Organization & People) Model, SI(S/W infrastructure) Model, and ST(S/W Technology) Model. As a model providing base information of the other models, Domain Model plays the role of tailoring guide controlling 4 sub models with domain information of each business unit. It reflects specific environmental information which business units act currently. The process evaluation model is developed through tailoring to fit the conditions of business units based on the IS15504 process model as an evaluation method of process capability. With the process evaluation model as the most top-level model, It was developed to be able to evaluate and analyze specific processes (much more important than one item in the process evaluation mode) which are not covered by the process evaluation model by developing organization competence evaluation, personnel, organization, infrastructure and technology as a separate model. Each model is divided into detail sub categories for more detail analysis and present status, gap analysis, and issues are analyzed for each item with the result of applied model, and it presents a practical method for the reinforcement of competence by extracting action item of model level.

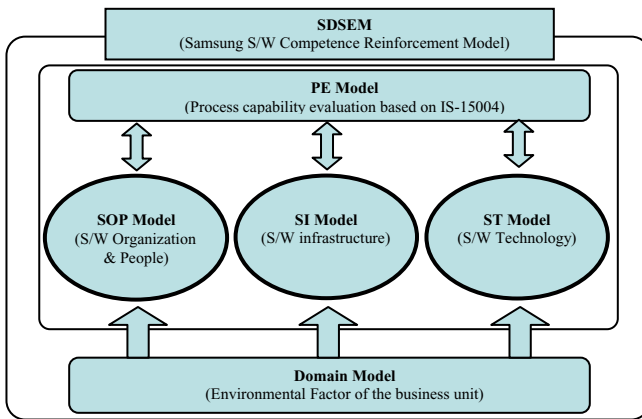


Fig. 2. Structure of the SDSEM

3.2 Process Evaluation Model (PE Model)

The S/W development competence can be evaluated by process capability, human resource, technology, infrastructure of an organization. The PE model complies with international standard (IS 15504) and CMMI. The target process area is tailored based on basic competence of business unit in the process area of a basic model.

Figure 3 shows a competence evaluation model developed in the corporate by tailoring the process area of IS15504. We have developed PE basic model to cover IS15504's minimal requirement of the level 2 capability. And the extended model, that is, the advanced version of basic model, was developed to cover SPICE level 3.

The extended model has 25 processes from 9 process groups including product release necessary in multi-site development, reuse process essential for the improvement of product quality and productivity, and 4 processes pertinent to quality. The extended model is applied to business unit already introduced business unit-wide standard process or reasonably equipped with competitiveness in software field, and the status of each business unit is determined by prior investigation.

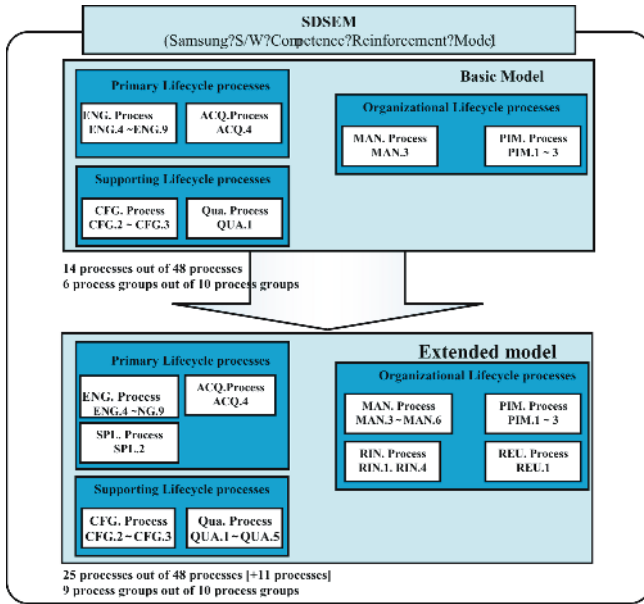


Fig. 3. Process area of the PE model

The process evaluation model in this study currently focuses on the area only for the software field, but we are going to develop more extended model with system level integration as an extended plan in a near future. For this, the Eng. Process of IS15504 will be extended and the introduction of CMMI may be considered.

3.3 SOP Model

The SOP (S/W Organization & People) model is for evaluating and analyzing an organization and its members, which are the entities performing business activities in an enterprise. Even the processes like IS15504 or CMMI already have process area for the two, a separate model is prepared and defined because the importance of this area is very high in the industry (it may be considered as one of the most important elements).

The first item, software developer against T&D total can be a raw data representing actual number of S/W personnel and it can be also used as an index indirectly representing the software portion and its importance level of a business unit. The second item, classification of staff by software development experience represents the personnel of a business unit by experience, so we can find out any problems related to the personnel

level distribution from experience level distribution. The third item is organizational hierarchy of personnel on software technology. The fourth item is organizational hierarchy. The fifth is S/W training system, and it's one of the important elements necessary to reinforce the development competence of an organization together with personnel acquisition. S/w training also depends on the S/W technology roadmap of the ST model (that is, this can be prepared once S/W technology roadmap is prepared). This analyzes the system to be used in acquiring important S/W technologies defined in the technology roadmap. The hierarchy of SOP model in Figure. 4.

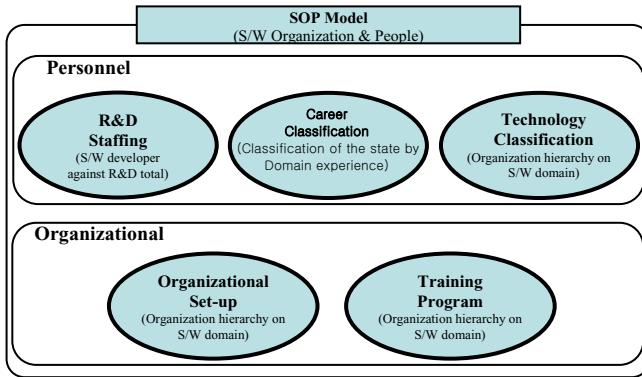


Fig. 4. Elements of the SOP model

3.4 ST Model

As a model to analyze the current technology level of an organization and define mid to long-term core technologies to acquire, the ST (S/W technology) model is for reinforcing the technology and personnel through interaction with the SOP model. The elements of the ST model includes software technology classification software technology roadmap, and reuse technology.

S/W technology classification is a framework for software technology of an organization, and it defines hierarchy and relationship of technologies necessary in preparing technology definition, personnel acquisition, and technical roadmap, The technical analysis by product can be achieved by separating and classifying embedded software according to product groups by the technologies being applied.

To keep pace with the speed of convergence and market changing, it is critical issues to build environment for reuse, create reusable assets which can be shared among development organizations. It assesses the reuse level of an organization and enables reuse policy establishment and technology acquisition suitable to each evaluation result. The evaluation of reuse level is performed by following organization level by reuse proposed by Ivar Jacobson[6]. The advantage of this evaluation will be prevention of duplicate development for defined technology group among organizations, mutual utilization of retained advantageous technologies, and reinforcement of development cooperation among organizations.

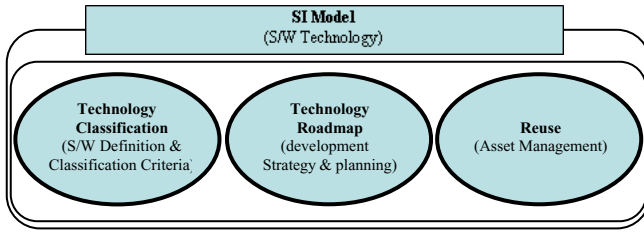


Fig. 5. Elements of the ST model

3.5 SI Model

The SI (S/W Infrastructure) Model evaluates the standardization and utilization status of tools necessary for S/W development, and it introduces issues to improve for the evaluation result. This model targets only tools used to develop S/W and focus on standardization of tools in business unit level and technical support for them. The first item lists all the tools currently being used in an organization for S/W development, and figures out the demand/supply of each tool. The second item, tool standardization status finds out the application level of company-wide standard tools through the analysis of acquisition and utilization status for the tools designated as company-wide standard tools. The third item, technical supporting finds out establishment and operation of environment for installation, maintenance and technical support to enable efficient utilization of tools in an organization as in figure 6.

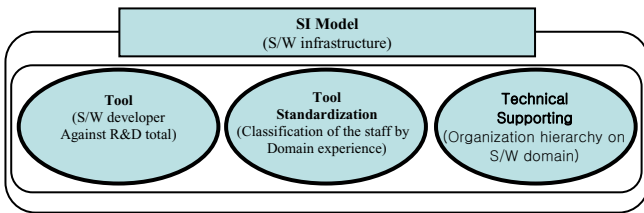


Fig. 6. Components of the SI model

4 Model Experimentation

This model was developed in terms of methodology to enable easy application in each business unit and to overcome the limitation of SPI Activity that uses only IS15504. It has been verified by applying the model to competence reinforcement activity of the business units.

4.1 Experimentation Process

The application result of SDSEM includes present status, gap analysis, issues and action items by each sub category. Among these, the action items define practical

actions for competence reinforcement after applying this model, and it should be accompanied by counteraction strategy for each action item, scheduling and resource allocation. Procedure for Applying SDSEM and its in/outputs are shown in Figure 7.

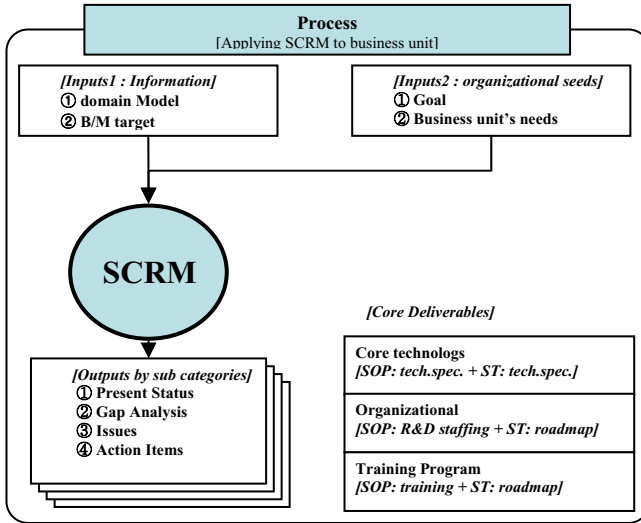


Fig. 7. SDSEM Process

4.2 Experimentation Results

It shows SDSEM is effective to our business domain by our goal to improve software development strength. To maintain acquired capability, we have conducted bi-annual corporate-wide assessment to check the conformance and maintenance of each business unit.

It indirectly shows that we have more powerful supporting capability to software development and that is one of results of applying SDSEM.

5 Conclusion

SDSEM is a framework for reinforcing the S/W competence based on IS15504 and CMMI. It is a practical solution to perform in industry. We introduced 5 sub models to abstract, simplify real world and they yield strategic direction and practical work items with specific plan in each. It is essential to make each work items to be manageable for accomplishing goal of applying SDSEM to the organization. It is also helpful for people who have responsibility to perform improvement job. SDSEM can be a formal guideline to be referred by business units and standard to regulate business units. SDSEM satisfy the minimal requirement at current situation in software domain and low level of maturity. But it is needed to be extended into product level system domain and to cover the higher level of maturity. It is also needed to be im-

proved to the statistical, quantitative analysis method for revealing the effectiveness of SDSEM. Correlations between sub models and their elements are to be analyzed and reflected to SDSEM itself.

References

1. KSPIICE (Korea Association of Software process Assessors), SPICE Assessment Report <http://kaspaa.org>, 2002~2004
2. ISO/IEC JTC1/SC7 15504: Information Technology-Software Process Assessment, ISO, ver.3.3, 1998
3. M. C. Paulk, M.D. Konrad, and S.M. Garcia, "CMM versus SPICE Architectures", Software Process Newsletter, No. 3, Spring 1995, pp. 7-11.
4. M. C. Paulk, "The Capability Maturity Model: Guidelines for Improving the Software Process", Addison-Wesley publishing, 1996.
5. ISO/IEC 15504-1:2004, "Information technology – Process assessment – Part 1: Concepts and vocabulary", 2004.
6. B. Boehm, Port D., "Escaping the Software Tar Pit: Model Clashes and How to Avoid Them", Software Engineering Notes, Association for Computing Machinery, pp. 36-48, 1999.
7. C. B. Seama, V. R. Basili, "Communication and Organization in Software Development: An empirical Study", IBM Systems Journal, 36(4), 1997.
8. I. Jacobson, M. Griss, P. Jonsson, "Software Reuse", Addison Wesley, pp. 15-24, 1997.
9. Tomer, A., Schach, S.R., The evolution tree: a maintenance-oriented software development model, Software Maintenance and Reengineering, 2000. Proceedings of the Fourth European, pp. 209-214, 2000.
10. Lattanze, A.J., Rosso-Llopert, M., Managing cyclical software development, Engineering and Technology Management, 1998. Pioneering New Technologies: Management Issues and Challenges in the Third Millennium. IEMC '98 Proceedings. International Conference on, pp. 62-70, 1998.

A Robust Routing Protocol by a Substitute Local Path in Ad Hoc Networks

Mary Wu¹, SangJoon Jung², Seunghwan Lee³, and Chonggun Kim^{1,*}

¹Dept of Computer Eng., Yeungnam Univ.,
214-1, Deadong, Kyongsan, Kyungbuk, 712-749 Korea
mrwu@yumail.ac.kr, cgkim@yu.ac.kr

²Dept. of General Education, Kyungil Univ.,
33, Buho-ri, Hayang-up, Kyongsan, Kyungbuk, 712-701 Korea
sjjung@kiu.ac.kr

³SoC R&D Center, System LSI division, Samsung Electronics Co.
seung1972@hotmail.com

Abstracts. Ad hoc wireless networks consist of mobile nodes in an area without any centralized access point or existing infrastructure. The network topology changes frequently due to nodes' migrations, signal interferences and power outages. One of the ad hoc network routing protocols is the on-demand routing protocol that establishes a route to a destination node only when it is required by a source node. The overhead of maintenance is low, but it is necessary to reestablish a new route when the routing path breaks down. Several recent papers have studied about ad hoc network routing protocols avoiding the disconnection of the existing route. Each active node on the routing path detects the danger of a link breakage to a downstream node and try to reestablish a new route. The node detects a link problem, try to find a substitute local path before the route breaks down. If the local reestablishment process fails, the route breaks down. In this paper, a robust routing protocol is proposed to increase the success rate of the local route reestablishment and enhance the communication delay performance. The results of computer simulation show that the proposed routing protocol increases system performance.

1 Introduction

Ad hoc network is the cooperative engagement of a collection of mobile nodes without the required intervention of any centralized access point or existing infrastructure, where all nodes are capable of moving and can be connected dynamically in an arbitrary manner. Nodes in a network function as routers, which discover and maintain routes to other nodes in the network. One important challenge in the design of ad hoc networks is the development of dynamic routing protocols that

* Correspondence author.

can efficiently find routes between two communicating nodes. The on-demand routing protocols establish a routing path to a destination node only when the source node has data to transmit[1-6]. AODV(ad hoc on-demand distance vector routing)[4], DSR(dynamic source routing)[5], TORA(temporary-ordered routing algorithm)[6], ABR(associativity based routing)[7], SSA(signal stability-based adaptive routing)[8] are categorized in this type. When a node requires a route to a destination, it initiates a route discovery procedure. This procedure is completed once a route is found or a possible route has been examined.

In order to obtain a stable routing path, ABR uses a cumulated number of beacon signals to measure the stability of neighboring nodes. The stability measurement factor is appended to search packets so that the destination node can construct a stable route according to the result of stability measurement. SSA uses the information of the signal strength at the link to choose a stable route. If there is a route failure due to host mobility, signal interference or power outage, it requires additional time and heavy traffic of reconfiguring the route from the source to the destination. ARMP(active route maintenance protocol)[9] tries to prevent disconnecting of the current route by monitoring the status of the signal strength and stability of the individual links. Two nodes of a weak link perform a local route reestablishment process to find a substitute local path before the routing path is broken. But, if the local route reestablishment process fails, the route may be broken. The success rate is important. In this paper, RRAODV(robust routing protocol of AODV) is proposed to increase the success rate of the local route reestablishment and enhance the route efficiency. RRAODV is on-demand routing protocol based on AODV. Simulation results demonstrate that the proposed RRAODV reduces the probability of local route breakage compared to that of the previous studies.

2 Overview of Local Recovery Methods of Route

2.1 ARMP

ARMP establishes a substitute partial route before the occurrence of route disconnection. When the state of a link is changed to an unstable state, the end nodes detect whether a route disconnection will be caused in the near future by the link status. One of the end nodes of this link is selected as an active node to establish a substitute local path.

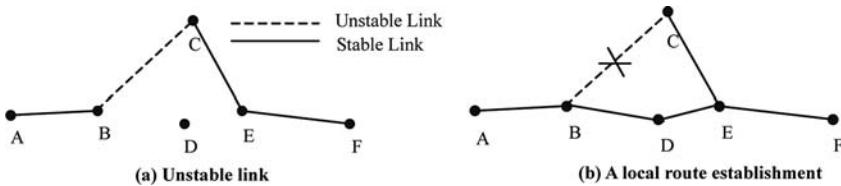


Fig. 1. A local recovery in ARMP

In fig.1, a neighbor node of node D will be selected as stepping node to connect node B with node E. Thus, a new local route can be established before the original route is disconnected. If the local route reestablishment process fails, the route is going to break.

2.2 Concepts of the Proposed Robust Routing Protocol

The proposed RRAODV is proposed to increase the success rate of the local route reestablishment. Active nodes which are nodes on the routing path monitor the signal strengths to the next node and check the stability of the link state. When the signal strength of a link is less than SS_{thr} (signal strength threshold of stable link state), the link is considered unstable. The node of an unstable link starts the reconstructing of a local routing path. The signal strengths of links on the ad hoc network are maintained in each node[9]. The reestablishment process has two steps. First, the node selects one of the downstream nodes in its transmission range as a new next node. In fig.2(a), the routing path has node S, A, B, C, E, F, G, D as active nodes at T_i . In fig.2(b), node E moves and node C has a weak link connection with the next node E at T_{i+1} . Node C starts to select a new next node. The downstream nodes of node C are nodes E, F, G, D. There is node F in the transmission range of node C at T_{i+1} . In fig.2(c), node C selects the downstream node F as the new next node.

If there isn't any downstream node in the transmission range of node C, the reestablishment process migrates to the second step. In fig.3(a), the routing path is S, A, B, C, E, F, G, D at T_i . Node E moves and node C has a weak connection to the

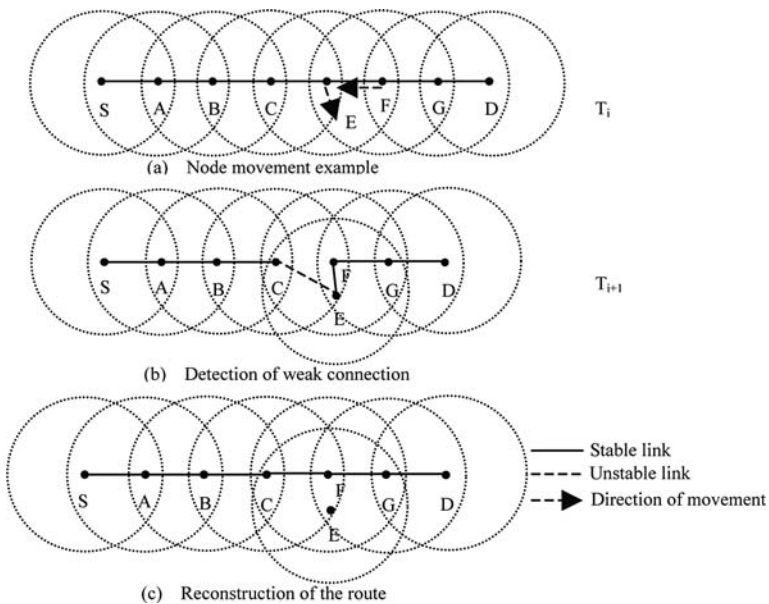


Fig. 2. Selection of one of the downstream nodes as a next hop node

next node E at T_{i+1} . In fig.3(b), there isn't any downstream node in the transmission range of node C at T_{i+1} . In this case, the first selection process is failed. In the second step, the node C selects one of the non-active neighbors which will become the new next node. The non-active node H is selected as the stepping node from node C, to connect to the downstream node F. As the result, node H is selected as the new next node of node C.

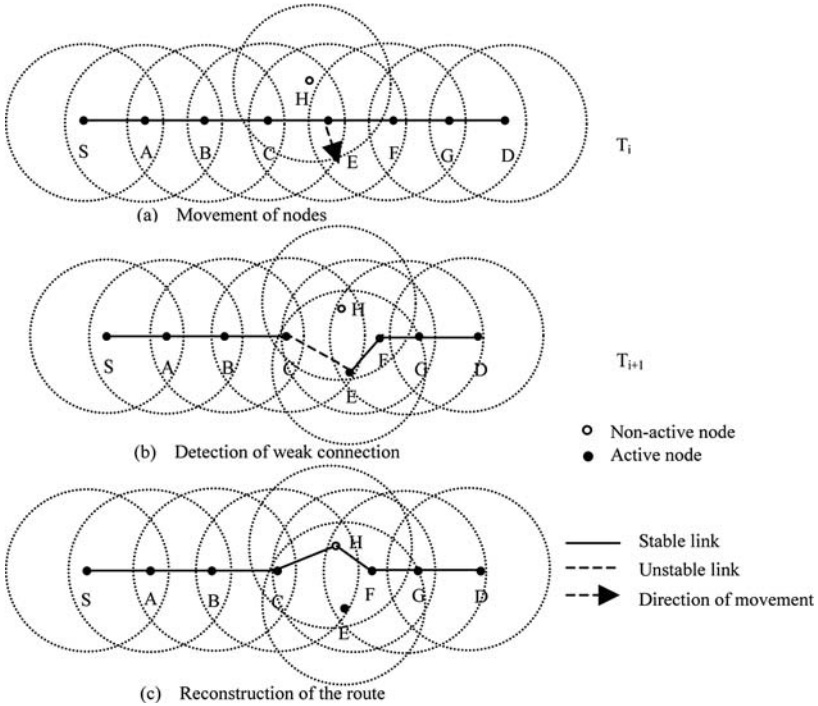


Fig. 3. Selection of one of the non-route neighbors which have connection to the downstream node

3 Operation of the Robust Routing Protocol

RRAODV mainly consists of three phases, the first phase is the ordinary route searching process. In the second phase, each node on the routing path monitors the link state to the next node. In the third phase, when a node has an unstable link to the next node, it requires a route reestablishment process.

3.1 Routing Table and Packet Format

The route discovery is started by broadcasting a RREQ(route request) packet to its neighbors. If a node which receives RREQ is an active node on the routing path to the destination or itself is the destination, the node transmits a RREP(route reply) back to

its previous neighbor which sent the RREQ. A node that received the RREP propagates the RREP toward the source. When each node receives the RREP, the node prepares routing informations and store it in the routing table. The routing table entry contains the following information[4]:

Destination, Next Hop, Number of hops, Sequence number for the destination, Active neighbors for this route, Expiration time for the route table entry.

In the proposed method, the information about downstream nodes is added to the routing table. In fig.4, node E receives a RREP from node F and records the information of downstream nodes G, D. In this case, node G is the first downstream node, and node D is the second downstream node. Table.1 is an example of the routing table on node E.

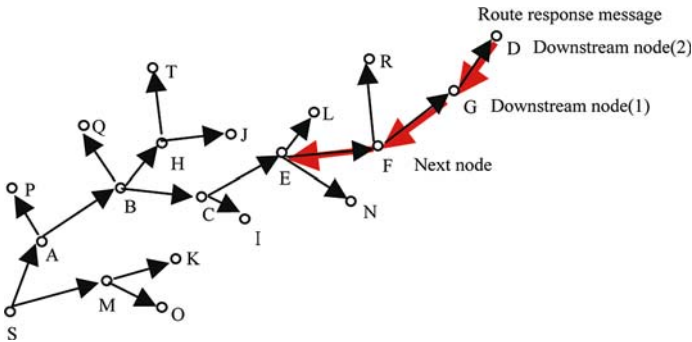


Fig. 4. RREP message transmission

Table 1. The routing table of the node E

Destination	Next hop	Hop count	Number of downstream	Downstream
D	F	2	1	G
			2	D

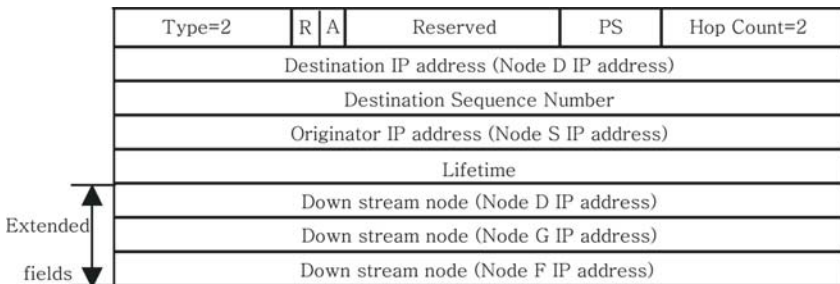


Fig. 5. The extended RREP message sent by node F to node E

When an intermediate node receives a RREP, it propagates the RREP toward to the source including its own IP address in a downstream node field of the RREP message. The extended RREP message format is shown in fig.5. Node F writes its own IP address to the downstream node field and propagates the message to node E.

3.2 Estimation of the Link Fault Using Link States

Local connectivity is confirmed by hello messages. If hello messages are not received from the next active node on the routing path during a certain interval, the node will send the notification of a link failure on AODV[4]. In the proposed method, each active node records the signal strength of hello messages received from neighbors. In fig.6, node E receives hello messages which have information about node F and the downstream nodes G, D.

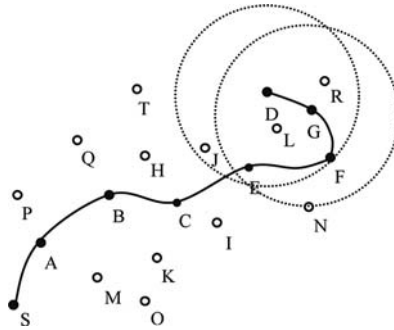


Fig. 6. An example of a routing path

Table 2. The table of the signal strength from the next node F(node E)

Time	Next node	Signal strength	Link state
HT _i	F	SS _{FE}	S(stable) or U(unstable)

Table 3. The table of signal strength from the downstream node G, D(node E)

Time	The number of downstream	Downstream	Signal strength
HT _i	1	G	SS _{GE}
HT _i	2	D	SS _{DE}

Nodes maintain the signal strength from the next node and the downstream nodes like table 2 and table 3. SS_{FE} is the signal strength which node E receives from node F.

Whenever each active node detects an unstable link state, a candidate node must be chosen to establish a substitute stable routing path. A link which the signal strength is less than SS_{thr} is considered unstable.

3.3 Selection of a New Next Node

All nodes have to manage a list of one-hop neighbors. By receiving a hello message from a new neighbor, or failing to receive consecutive hello messages from previous neighbors, a node can detect that the local connectivity has changed[4].

The selection of a new next node has two steps. First, the node has an unstable link to the next node selects one of the downstream nodes in its transmission range as a new next node. If there isn't any downstream node in the transmission range, as the second step, the current node selects one of non-active neighbors which can connect to one of the downstream nodes as the new next node. If the second step also fail, the current node transmits the RERR(route error) message to the source node and it restarts the route discovery process as AODV. In the first local recovery process, when downstream nodes are available as next node candidate, then the nearest node to the destination node have to be selected as the next node.

In table.3, if SS_{DE} is stronger than SS_{thr} , node D is chosen as a new next node. If not, node E checks SS_{GE} . If SS_{GE} is stronger than SS_{thr} , node G is chosen as a new next node. If there isn't any node has that the signal strength is stronger than SS_{thr} in the table, node E initiates the second selection process.

In the second selection process, the node broadcasts a SSRQ(signal strength request) message to select a stable non-active node which will work as a stepping node to connect to one of the downstream nodes. In fig.7(a), node E broadcasts a SSRQ packet, then one-hop-neighbor nodes I, J, L, N receive it. The SSRQ includes the information of the downstream nodes and the signal strength in fig.8(a). The non-active nodes I, J, L, N check the information of the downstream nodes. If they have some information from the downstream nodes, they reply SSRP(signal strength reply) packets including the signal strength information to node E. There is node L in the transmission range of the downstream node G and the next node F in fig.7(b). Therefore, node L has managed the signal strength of node G and node F. When node L receives the SSRQ, it replies the SSRP to node E including the signal strength of nodes G and node F.

The signal strength message format is shown in fig.8. It contains a type value of 5. The code of SSRQ message value is 1, and SSRP is 2. Node E receives SSRP packets from nodes L, N and selects node L as a new next. Node G is also selected as the next of the next node on the routing path.

Node E selects one of non-active neighbors with a stable links to the downstream nodes on the routing path. The selected node must have the sufficient signal strength both the previous node and the downstream node.

In the table 4, the one-hop neighbors of node E are nodes I, J, L, N. There isn't any neighbor within the transmission range of the 2nd downstream node D. Therefore, the values of SS from the 2nd downstream field are all 0s. There are nodes L, N in the transmission range of the 1st downstream node G. The signal strength of node L from the 1st downstream node G is SS_{GL} , and that of node N is SS_{GN} . If SS_{GL} is stronger than SS_{thr} , the node is selected. If SS_{GL} is weaker than SS_{thr} , and SS_{FL} is stronger than SS_{thr} , node L is selected. Both SS_{GL} and SS_{FL} are weaker than SS_{thr} , then node E transmits a RERR message to the source node.

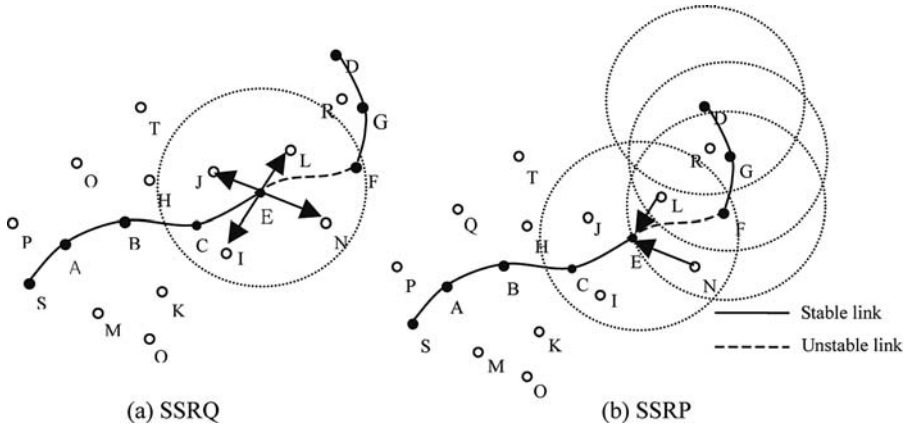


Fig. 7. Gathering of the signal strength message

Type =5	Code =1	Reserved
Request node address = E		
Signal strength		
Next node address = F		
Signal strength		
1st downstream node address = G		
Signal strength		
2nd downstream node address = D		
Signal strength		

(a) SSRQ

(a) SSRQ

(a) SSRQ

Type =5	Code =2	Reserved
Request node address = E		
Signal strength = SS_{EL}		
Next node address = F		
Signal strength = SS_{GL}		
1st downstream node address = G		
Signal strength = SS_{GL}		
2nd downstream node address = D		
Signal strength		

(b) SSRP

(b) SSRP

(b) SSRP

Fig. 8. The format of the signal strength message

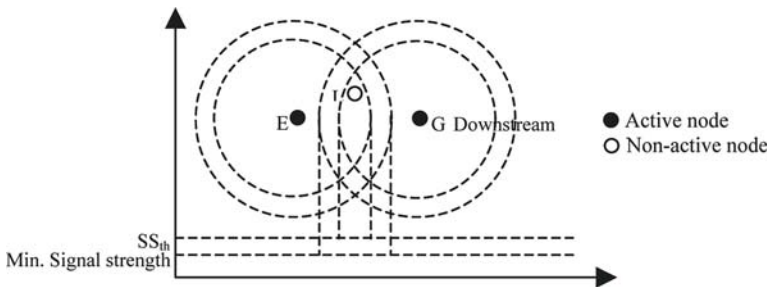


Fig. 9. Selection of one of the neighbors which can connect to the downstream node

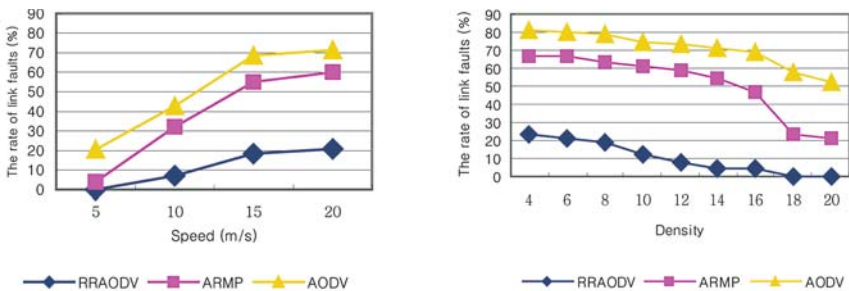
Table 4. The table of the signal strength

One-hop neighbor	SS from node E	SS from next node F	SS from 1 st downstream G	SS from 2 nd downstream D
I	SS _{EI}	0	0	0
J	SS _{EJ}	0	0	0
L	SS _{EL}	SS _{FL}	SS _{GL}	0
N	SS _{EN}	SS _{FN}	SS _{GN}	0

4 Experiments and the Results

A computer simulation is derived to evaluate performance, the rate of link faults, the success rate of the selection process, the delay of RRAODV, ARMP, and AODV. The MANET size is 1000 × 1000 units. Nodes move at speeds 5, 10, 15, 20, 25 or 30 units per second. The transmission range is 250 units. The SS_{thr} of a stable link is set to 10 units. The fig.10(a) shows the rate of link faults, as the node speed is varied and the node density is set to 16. The results show that the rate of link faults increases, as the speed increases and the rate of link faults of RRAODV is less than those of ARMP or AODV. The fig.10(b) shows the rate of link faults, as the node density is varied and the node speed is set to 15(m/s). The rate of link faults of RRAODV is less than that of ARMP or AODV. The results also show that the rate of link fault decreases as the density increases.

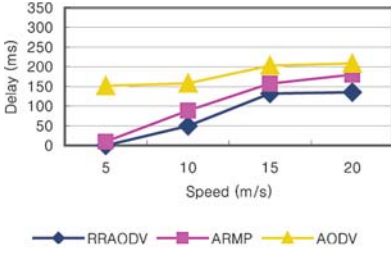
The fig.11(a) shows the delay as the node speed is varied. The node density is set to 16 and SS_{thr} is set to 10. The delay increases as the node speed increases. The delay of RRAODV is less than that of ARMP or AODV. The fig.11(b) shows the delay as the node density is varied. The node speed is set to 15(m/s) and SS_{thr} is set to 10. The delay decreases as the node density increases. The delay of RRAODV is also less than that of ARMP or AODV.



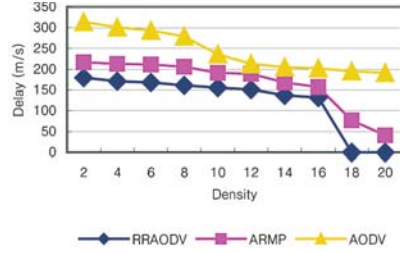
(a) The rate of link faults depending on speed

(b) The rate of link faults depending on density

Fig. 10. Comparison of the rate of link faults in the RRAODV, the ARMP, and the AODV

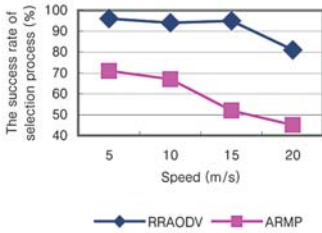


(a) The delay depending on speed

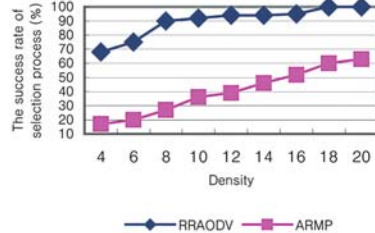


(b) The delay depending on density

Fig. 11. Comparison of the delay in the RRAODV, the ARMP, and the AODV



(a) The success rate of the selecting process depending on speed



(b) The success rate of the selecting process depending on density

Fig. 12. The success rate of the substitute local routing path selecting process in the RRAODV and the ARMP

The fig.12(a) shows the success rate of the local substitute routing path selecting process of RRAODV and ARMP as the node speed is varied. The node density is set to 16 and SS_{thr} is set to 10. The fig.12(b) shows the success rate of the local substitute routing path selecting process of RRAODV and ARMP as node density is varied and the node speed is set to 15(m/s). The results show that the success rate of RRAODV is larger than that of ARMP.

5 Conclusions

We proposed RRAODV routing protocol for preventing the current route from disconnection. In RRAODV, by extending the range of node selection, the success rate of the substitute local routing path selection process is raised. It also reduces the probability of route breakage than that of ARMP and AODV. By the computer simulation, the local recovery delay RRAODV is apparently lower than that of AODV and the success rate to keep the route to the destination increases. A practical implementation of the protocol is also an essential future work.

References

1. J.J. Garcia-Luna-Aceves, M. Mosko, and C. Perkins, "A New Approach to On-Demand Loop-Free Routing in Ad Hoc Networks," Proc. Twenty-Second ACM Symposium on Principles of Distributed Computing (PODC 2003), July 13-16, 2003.
2. J. Raju and J.J. Garcia-Luna-Aceves, "Efficient On-Demand Routing Using Source-Tracing in Wireless Networks," Proc. IEEE Global Telecommunications Conference (GLOBECOM), Nov. 27-Dec. 1, 2000.
3. C. E. Perkins and E. M. Royer, "Ad-hoc On-Demand Distance Vector Routing," in Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications, New Orleans, LA, pp. 90-100, Feb. 1999.
4. C. Perkins, E. Belding-Royer, S. Das, "RFC 3561 - Ad hoc On-Demand Distance Vector (AODV) Routing," 2003.
5. D. B. Johnson and D. A. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Networks," in Mobile computing, T. Imielinski and H. Korth, Eds. Kluwer Academic Publishers, pp. 153-181, 1996.
6. V. D. Park and M. S. Corson, "A Highly Adaptive Distributed Routing Algorithm for Mobile wireless Networks," in Proceedings of IEEE INFOCOM'97, pp. 1405-1413, Apr. 1997.
7. C.-K. Toh, "Associativity-Based Routing For Ad-Hoc Mobile Networks," Journal on Wireless Personal Communications, vol.4, First Quarter, 1997.
8. R. Dube, et al, "Signal Stability based Adaptive Routing(SSA) for Ad Hoc Mobile Networks," IEEE Personal Communication Magazine, Feb. 1997.
9. Chih-Yung Chang, Shin-Chih Tu, "Active route-maintenance protocol for signal-based communication path in ad hoc networks", Journal of Network and Computer Applications (2002), Vol 25, 161-177.
10. Mary Wu, YoungJun Choi, Younseok Jung, KyungSoo Lim, Chonggun Kim, "A Calculation Method of Handoff Rate using Speed and Direction of Mobile Stations in Cellular Networks, " Journal of Kiss, Information Networking Vol. 29, Num. 4, 2002. 8.

Power Efficient Wireless LAN Using 16-State Trellis-Coded Modulation for Infrared Communications

Hae Geun Kim

School of Computer and Information Communication,
Catholic University of Daegu,
330 Kumrak-ri, Hayang-up, Kyungsan-si, 712-702, Korea
kimhg@cu.ac.kr

Abstract. Optical wireless communication employing 16-state trellis-coded multiple-subcarrier modulation with a fixed bias is introduced, where 4-dimensional vectors having the largest Euclidean distance for 32 signal points are used. After combining coding with the 4-dimensional modulation, the proposed system improves the power and the bandwidth efficiency, significantly. The performance evaluation results are that the normalized power and the bandwidth requirements are reduced up to 6.1 dB and 5.8 dB compared to those of the block codes using QPSK, respectively. The proposed system needs much less power and bandwidth than the counterparts transmitting the same bit rates for optical wireless connection.

1 Introduction

A Wireless Local Area Network (WLAN) provides over the air interface standardized in IEEE 802.11 that includes two different technologies based on infrared and radio frequency (RF) for the physical layer [1]. Infrared wireless local access links offers a virtually unlimited bandwidth at low cost because optical sources and receivers are capable of high-speed operation.

When an optical WLAN is used for indoor use there are two major challenges that are the scattering of the light by the interior of room and power efficiency. To overcome those issues, Multiple-Subcarrier Modulation (MSM) systems with Intensity Modulation / Direct Detection (IM/DD) in optical wireless communications are popularly researched [2], [3], [4]. IM/DD MSM systems are attractive not only for minimizing inter-symbol interference (ISI) on multi-path channels between narrowband subscribers, but also for providing immunity to ambient light inducing in an infrared receiver. Yet, the efficiency of average optical power intensity is poor as the larger number of subcarriers in an MSM system.

In this paper, the MSM for infrared wireless communications using 16-state trellis-coded modulation (TCM) [7] is introduced where the newly generated 4-Dimensional (4-D) vectors for 32 signal symbols with the maximized minimum distances are employed. The block coder of the proposed system including a 4/5 convolutional

encoder and an impulse generator maps the information bits to be generated to the symbol amplitudes modulated on to the subcarriers. The computer derived 4-D vectors for 32 symbols are symmetrically constructed on the surface of a 4-D sphere. Also the fixed bias is used for all symbols so that the power used for each symbol is constant and equals the average transmitted power. In the proposed system, one symbol is transmitted with 4 orthogonal subcarrier signals, while one symbol is transmitted with one subcarrier in conventional On-OFF Keying (OOK) and with two subcarriers in Quadrature Phase Shift Keying (QPSK). The performance evaluation for each scheme is carried out, and the power and bandwidth requirements of the proposed system are compared with that of the three block coders using QPSK scheme as counterparts.

2 4-D IM/DD MSM System in Optical Channel

Fig. 1 depicts the transmitter and receiver design used in the proposed MSM transmission scheme with 4-D orthogonal modulation where the transmitter in Fig. 1 (a) transmits Nk information bits during each symbol interval of duration T . In the block coder of each 4-D MSM in Fig. 1 (b), k input bits are encoded by $k/k+1$ convolutional encoder. The $k+1$ encoder output bits are transformed into one of M symbols, A_i , where $M = 2^{k+1}$ and, $i = 0, \dots, M$, then, each symbol A_i is mapped to a corresponding vector of 4-D symbol amplitudes, $\mathbf{a}_i = (a_{i1}, a_{i2}, a_{i3}, a_{i4})$. In the proposed system, $k = 4$ is chosen, so 32 symbols are generated for each 4-D MSM.

Since the electrical MSM signal $s(t)$ can be negative or positive, a baseband dc bias $b(t)$ must be added. After optical modulation using a LED or a LD, the MSM optical output in Fig. 1 (a) can be expressed as $x(t) = A[s(t) + b(t)] \geq 0$ where A is a nonnegative scale factor. The average optical power is $P = AE[s(t)] + AE[b(t)]$ where $E[x]$ represents an expected value of x .

If the subcarrier frequency $\omega_n = n(2\pi/T)$ and $g(t)$ is used in the MSM, $E[s(t)]$ is always 0 and the optical power P only depend on $b(t)$. Hence the average optical power can be given by $P = AE[b(t)]$. When the bias signal is properly chosen, the average power requirement of the MSM system can be decreased.

For fixed bias, the bias has the same magnitude with the smallest allowable value of electrical MSM signal $s(t)$ given by $b_0 = -\min_t s(t)$. For time-varying bias, the bias has the smallest allowable symbol-by-symbol value. In general, the average optical power with time-varying bias is smaller than that with fixed bias. On the other hand,

the system using fixed bias is simple and easy to implementation. In proposed system, fixed bias is used and has the bias value of $0.5s(t)$ that ensures nonnegative MSM signal to modulate an optical signal.

In the receiver shown in Fig. 1(a), the Gaussian noise is detected after the photo detector (PD). The converted electrical signal is divided into N 4-D receivers and demodulated with 4 orthogonal signals. In each 4-D receiver, 4 hard decision devices are used in obtaining a 4-D vector of detected symbol amplitudes $\mathbf{a}_i = (\hat{a}_{i1}, \hat{a}_{i2}, \hat{a}_{i3}, \hat{a}_{i4})$. The MLSD (maximum likelihood sequence detection) decoder regenerates the detected information bits $\hat{x}_1, \dots, \hat{x}_k$.

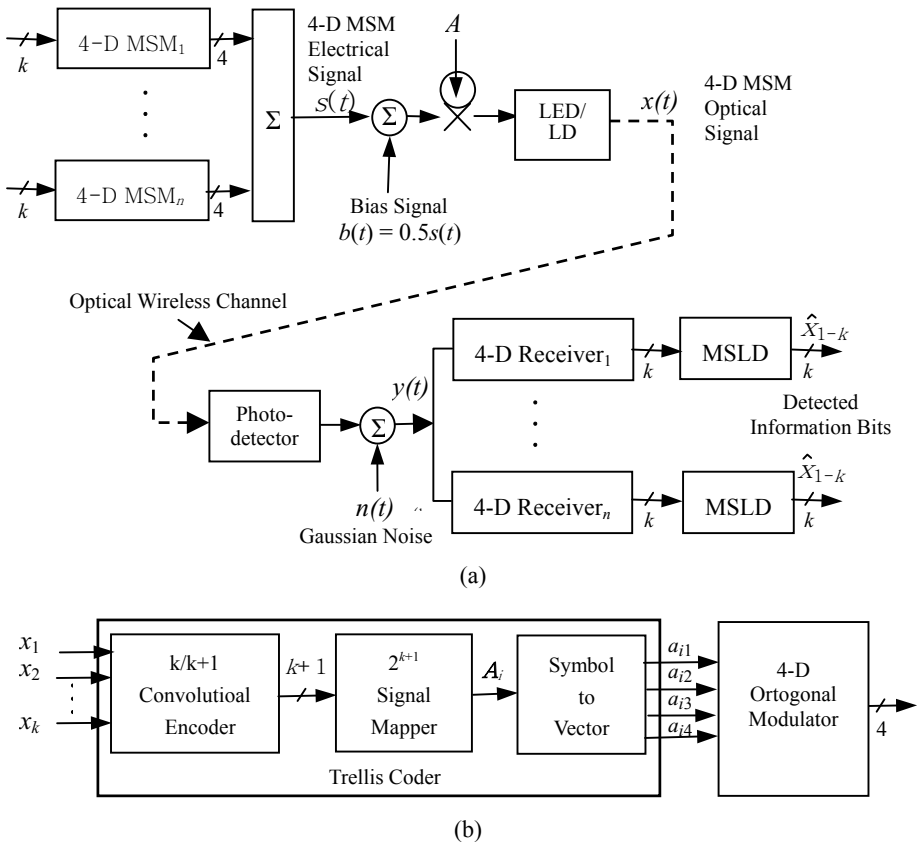


Fig. 1. MSM (Multiple-Subcarrier Modulation) system: (a) transmitter and receiver with 4-D orthogonal modulation scheme where $n = 1, 2, \dots, N$, and (b) 4-D trellis coder with orthogonal modulator. In the proposed system, the number of input bits $k = 4$ is used.

3 Generation of 4-D Vectors for the Proposed System

The error probability is almost entirely contributed by the nearest pair of signal points in an M -ary PSK scheme. For the case of high signal-to-noise ratio, we can choose a law of force expression, which can be given by [5].

$$F_{ik} = C(k_0)e^{-|d_{ik}|^2/4k_0/T} d_{ik}/|d_{ik}| \quad (1)$$

where F_{ik} is the force between particle i and k , k_0 is the bandwidth of noise, and d_{ik} is the vector from particle i to particle k .

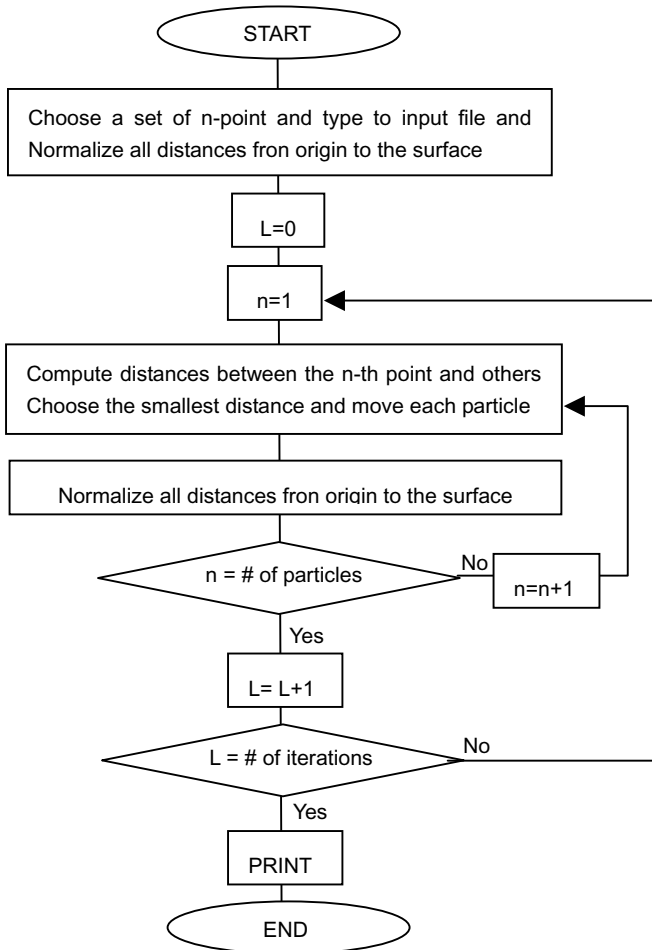


Fig. 2. The flowchart for the Problem

The program operates as the flowchart shown in Fig. 2 and has run for 32 points on the surface of a 4-D sphere. The results are shown in Table 1 where the symbols A_i are corresponding to output of 2^{k+1} signal mapper in Fig.1 (a) and the symmetric 4-D vectors are listed in an ascending order of the squared distances. The derived squared distance d^2 is 0 for A_1 itself, 0.8086 for A_1 to A_2 , 3.8666 for A_1 to A_{32} , and so forth and the magnitude of each vector $|A_i| = \sqrt{a_{i1}^2 + a_{i2}^2 + a_{i3}^2 + a_{i4}^2} = 1$ so that this block code is symmetry and each 4-D symbol has the same energy as any other symbol with the minimum Euclidean distance. The signal number of each signal A_i is given in order to obtain the maximum subset distance for set partitioning of trellis coding that will be explained in Sect. 4.1. In the proposed system, 4-D vector, $\mathbf{a}_j = (a_{j1}, a_{j2}, a_{j3}, a_{j4})$, in Table 1 is used in transmitting 4 information bits for the block coder in Fig. 1(a).

4 Design of 16 State TCM

TCM has been developed actively, since the publication by Ungerboeck [6] appeared as a combined coding with modulation technique for digital transmission. The main advantage of TCM is the significant coding gain achieved over conventional uncoded modulation on the severely bandlimited additive white Gaussian noise channel while maintaining the same bandwidth and transmitted power.

In the proposed system, the number of input bits k is 4, so a rate 4/5 systematic convolutional encoder with feedback is used for the 16-state TCM schemes. In the trellis diagram, one of 16 possible signals listed in Table 2 is diverged from a trellis node because the number of input bit is four and one uncoded input bit x_1 provides parallel transitions of 2 signals.

In order to determine the minimum distance of a TCM scheme, set partitioning of the signals is necessary. 16 subsets which have different minimum squared distances in Table 2 are formed. Two signals in each subset are associated with parallel transitions, so that the minimum subset distance should be maximized. In order to obtain the maximum subset distances, after comparing all the distances between signal points, 16 subsets are chosen as shown in Table 2. Since the squared distance of each subset is greater than or equal to 3.459, the minimum squared distance associated with parallel transitions can be 3.459.

Without loss of generality, we choose the upper all-zero path to be correct and the lower path to be incorrect path that represents an error event as shown in Fig. 3. From the trellis in Fig. 3, the minimum squared distance of 16-state TCM scheme is $d_m^2 = 0.809 \times 4 = 3.236$ because the distances associated with parallel transitions which are larger than this value.

Table 1. Computer derived 32-point vectors with maximized minimum squared distance on the surface of 4-D sphere

A_i	a_{i1}	a_{i2}	a_{i3}	a_{i4}	d^2 to A_1	Signal #
A_1	-0.362787	0.535530	0.077936	-0.758630	0.0000	0
A_2	-0.427980	-0.272260	0.466066	-0.724907	0.8086	1
A_3	0.451956	0.363509	-0.261169	-0.771614	0.8086	2
A_4	0.294727	0.241430	0.612178	-0.692882	0.8086	3
A_5	-0.691247	0.688544	-0.219103	0.008850	0.8086	4
A_6	-0.259356	0.371163	-0.753521	-0.476633	0.8086	5
A_7	-0.856946	-0.067417	-0.245302	-0.448248	0.8086	6
A_8	-0.116650	0.789680	0.591972	-0.111212	0.8086	7
A_9	0.156285	0.940721	-0.261711	-0.143747	0.9209	8
A_{10}	-0.181843	-0.018111	0.980102	-0.077489	1.6171	9
A_{11}	-0.701912	0.319009	0.513283	0.376952	1.6410	10
A_{12}	0.379311	-0.536620	0.171329	-0.734036	1.7095	11
A_{13}	0.773160	0.565435	0.271390	-0.094097	1.7703	12
A_{14}	0.283769	-0.376914	-0.688978	-0.550200	1.8822	26
A_{15}	-0.354732	-0.810617	-0.185695	-0.427296	1.9915	13
A_{16}	-0.787337	-0.502078	0.352972	0.058555	2.0003	28
A_{17}	-0.109646	0.770608	0.073873	0.623445	2.0295	27
A_{18}	0.527609	0.319783	-0.785662	-0.045853	2.0932	25
A_{19}	-0.407299	-0.368404	-0.835673	0.006029	2.2385	14
A_{20}	-0.815783	-0.052073	-0.288707	0.498432	2.2651	15
A_{21}	-0.192989	0.418606	-0.741856	0.487005	2.2662	19
A_{22}	0.935207	-0.175746	-0.187522	-0.243593	2.5264	31
A_{23}	0.333593	0.338596	0.716349	0.510795	2.5427	30
A_{24}	0.707852	-0.243788	0.660209	-0.060314	2.5803	20
A_{25}	0.014913	-0.810990	0.579961	-0.075618	2.6743	24
A_{26}	0.646590	0.374489	-0.208430	0.631059	3.0580	29
A_{27}	-0.200860	-0.359605	0.616925	0.670629	3.1608	21
A_{28}	0.430985	-0.877479	-0.210394	0.004127	3.2916	23
A_{29}	-0.289272	-0.774417	-0.168628	0.536810	3.4603	18
A_{30}	-0.075166	-0.021930	-0.143978	0.986478	3.4881	17
A_{31}	0.322350	-0.347136	-0.713213	0.516637	3.5007	16
A_{32}	0.581397	-0.433300	0.179701	0.664782	3.8666	22

Table 2. The Output Signal Sequences Diverged from Current States for 16 State TCM Scheme

Current state	Output Signals
0	(0, 16) (8, 24) (4, 20) (12, 28) (2, 18) (10, 26) (6, 22) (14, 30)
1	(1, 17) (9, 25) (5, 21) (13, 29) (3, 19) (11, 27) (7, 23) (15, 31)
2	(8, 24) (0, 16) (12, 28) (4, 20) (10, 26) (2, 18) (14, 30) (6, 22)
3	(9, 25) (1, 17) (13, 29) (5, 21) (11, 27) (3, 19) (15, 31) (7, 23)
4	(4, 20) (12, 28) (0, 16) (8, 24) (6, 22) (14, 30) (2, 18) (10, 26)
5	(5, 21) (13, 29) (1, 17) (9, 25) (7, 23) (15, 31) (3, 19) (11, 27)
6	(12, 28) (4, 20) (8, 24) (0, 16) (14, 30) (6, 22) (10, 26) (2, 18)
7	(13, 29) (5, 21) (9, 25) (1, 17) (15, 31) (7, 23) (11, 27) (3, 19)
8	(2, 18) (10, 26) (6, 22) (14, 30) (0, 16) (8, 24) (4, 20) (12, 28)
9	(3, 19) (11, 27) (7, 23) (15, 31) (1, 17) (9, 25) (5, 21) (13, 29)
10	(10, 26) (2, 18) (14, 30) (6, 22) (8, 24) (0, 16) (12, 28) (4, 20)
11	(11, 27) (3, 19) (15, 31) (7, 23) (9, 25) (1, 17) (13, 29) (5, 21)
12	(6, 22) (14, 30) (2, 18) (10, 26) (4, 20) (12, 28) (0, 16) (8, 24)
13	(7, 23) (15, 31) (3, 19) (11, 27) (5, 21) (13, 29) (1, 17) (9, 25)
14	(14, 30) (6, 22) (10, 26) (2, 18) (12, 28) (4, 20) (8, 24) (0, 16)
15	(15, 31) (7, 23) (11, 27) (3, 19) (13, 29) (5, 21) (9, 25) (1, 17)

5 Block Codes for the MSM System Using QPSK Scheme

For MSM systems, several block codes employing QPSK such as normal block code, reserved-subcarrier block code, and minimum-power block code to improve the power efficiency of an MSM optical communication system have been introduced [2].

Table 3. Set Partitioning of 32 Signals into 16 Subsets with $d^2 < 3.459$

Subset	Elements of Signals	Squared distance of subset
D ₀	0, 16	3.500
D ₁	1, 17	3.459
D ₂	2, 18	3.565
D ₃	3, 19	3.459
D ₄	4, 20	3.605
D ₅	5, 21	3.732
D ₆	6, 22	3.622
D ₇	7, 23	3.736
D ₈	8, 24	3.802
D ₉	9, 25	3.736
D ₁₀	10, 26	3.761
D ₁₁	11, 27	3.800
D ₁₂	12, 28	3.605
D ₁₃	13, 29	3.528
D ₁₄	14, 30	3.712
D ₁₅	15, 31	3.642

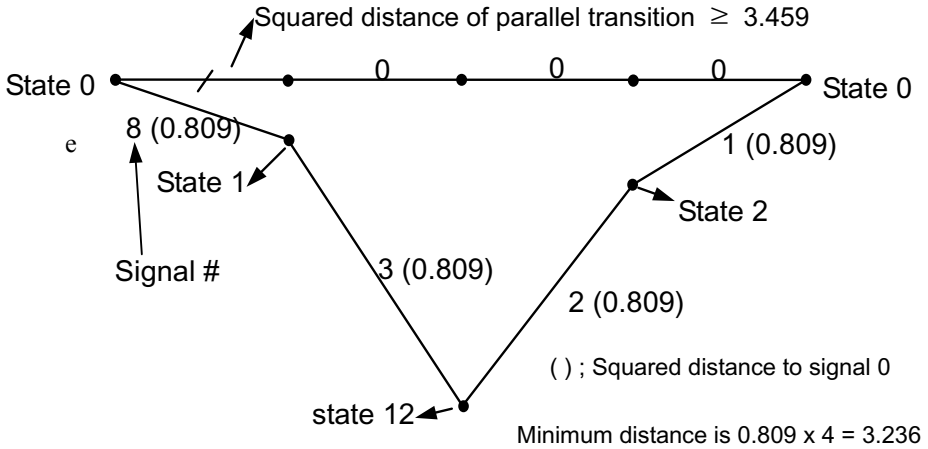


Fig. 3. The trellis to calculate the minimum distance of the 16 state TCM scheme

Under the normal block coder, all N subcarriers are used for transmission of information bit where the number of input bits $k = 2N$, and the number of symbol $M = 2^{2N}$ for QPSK. Each information bit can be mapped independently to the corresponding symbol amplitudes. At the receiver, each detected symbol amplitude can be mapped independently to information bits.

Under the reserved-subcarrier block code, L subcarriers are reserved for minimizing the average optical power P . Hence, the number of input bits $k = 2(N-L)$, and the number of symbol $M = 2^{2(N-L)}$ for QPSK. An information bit vector is encoded by freely choosing the symbol amplitudes on the reserved subcarriers.

Under the minimum-power block code, no fixed set of subcarrier is reserved, but $L > 0$ subcarriers are reserved for the minimum value of the average optical power P . Also, the number of input bits $k = 2(N-L)$, and the number of symbol $M = 2^{2(N-L)}$ for QPSK and the average optical power always lower bounds the average optical power requirement.

6 Performance Evaluation

The power and bandwidth requirements of the proposed system are compared with that of the three block coders using QPSK scheme described in Sect. 5 as counterparts. For the MSM system with M -ary PSK including the proposed system and QPSK schemes, the signal is composed of a sum of modulated sinusoids so that the bit error probability can be [7]

$$P_b = Q\left(\sqrt{r^2 A^2 T / 2N_0}\right) \quad (2)$$

where T represents the rectangular pulse duration, r is the responsivity of photodetector in (1) and $Q(x)$ is the Gaussian error integral, and A is a nonnegative

scale factor. In the proposed system, one symbol is transmitted with 4 orthogonal subcarrier signals, while one symbol is transmitted with two subcarriers in QPSK. When we consider the number of input bits and subcarriers, N 4-D MSM system is equivalent to $2N \times$ QPSK as described in Fig. 5. The error probability of each scheme is easily calculated with the minimum Euclidean distance. We set the required bit error probability to $P_b=10^{-6}$ [3].

Fig. 4 represents the numerical results of the normalized power requirement in optical dB versus the number of subcarriers with fixed bias for the proposed system and three block codes employing QPSK. Here, the normalized power requirements for three block codes employing QPSK are the results from [2] for fixed bias. The normalized power requirement of the proposed 4-D MSM is reduced up to 6.1 dB compared to those of above three block codes for QPSK scheme when the number of subcarriers is 4. For the number of subcarriers is 8, 12, and 16, the power requirement is reduced to 4.1 ~ 5.8 dB, 3.2 ~ 5.8 dB, and 2.8 ~ 5.8 dB, respectively. The power requirement in terms of the bandwidth requirement is also compared to measure the electrical bandwidth efficiency of the optical signal. Fig. 5 represents the normalized power requirement in optical dB versus the normalized bandwidth requirement with fixed bias for above three block codes employing QPSK and the proposed scheme. In the range of 1.125 ~ 1.5 of the normalized bandwidth requirement, the proposed system reduces up to 5.8 dB in bandwidth requirement. The proposed system has a large minimum value of 0.5s(t) and a large squared minimum distance of 3.236, so that the required dc bias is minimized and the error rate performance is improved. Hence, for wireless LAN using infrared communication, the proposed system can be much more efficient than the block codes using QPSK scheme for transmission of high-speed data with low power via the narrowband channel.

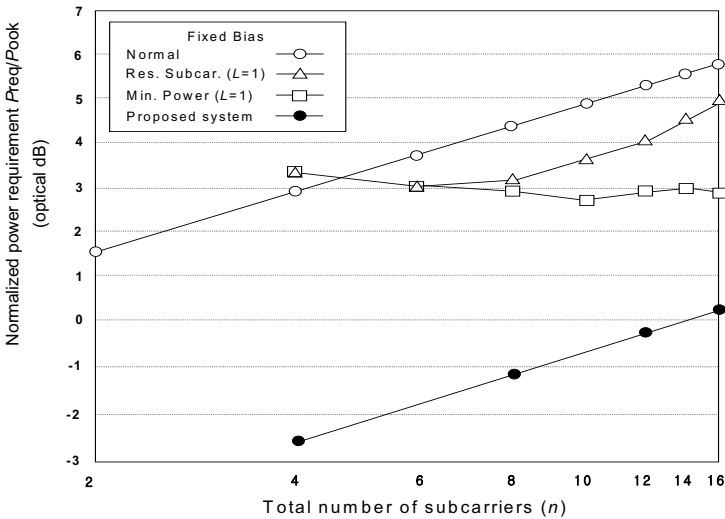


Fig. 4. Normalized power requirement versus total number of subcarriers for Normal QPSK, Res. Subcarrier, Min. Power, and the proposed 4-D MSM system

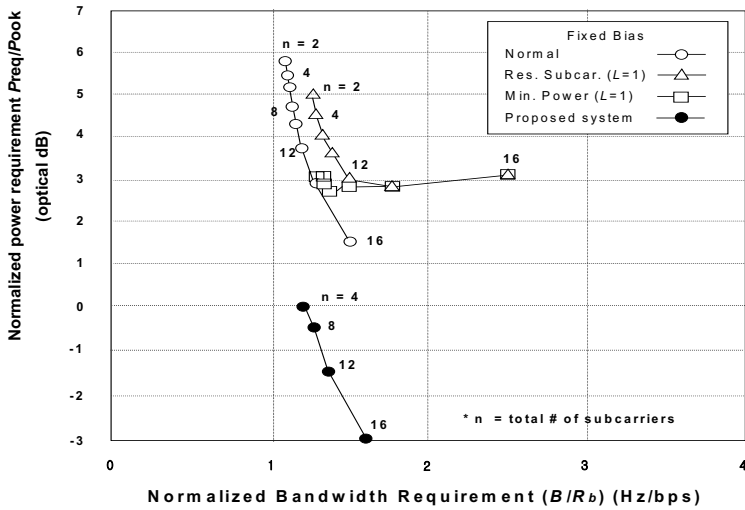


Fig. 5. Normalized power requirement versus normalized bandwidth requirement for Normal QPSK, Res. Subcarrier, Min. Power, and the proposed 4-D MSM system. Each number denotes total number of subcarriers.

7 Conclusions

This paper has described the basic principles and characteristics of multiple subcarrier modulation techniques in infrared communication for wireless LAN. The proposed system with the TCM scheme has a large squared minimum distance, so that the required dc bias is minimized and the error rate performance is improved. And, the optimization of signal waveform technique is used in deriving 4-D vectors for 32 points on the surface of Euclidean sphere having minimum distances between signal points. The 4-D MSM with fixed bias for optical wireless system using 4-D block coder improves the power and bandwidth efficiency, significantly. Hence, the proposed system needs much less power and bandwidth than the counterparts transmitting the same bit rates for optical wireless connection.

References

1. IEEE standard, <http://standards.ieee.org/getieee802/802.11.html>
2. Ohtsuki T.: Multiple-Subcarrier Modulation in Optical Wireless Communications. Vol. 3. IEEE Commun. Mag. (2003) 74-79
3. Kahn J., Berry J.: Wireless Infrared Communications. Vol. 2. Proc. IEEE (1997) 265-298
4. Teramoto S., Ohtsuki T.: Multiple-subcarrier Optical Communication System with Subcarrier Signal Point Sequence. IEEE GLOBECOM 2002

5. Lachs G.: Optimization of Signal Waveforms. Vol. 4. IEEE Trans. Information Theory (1963) 95-97
6. Ungerboeck G.: Channel Coding with Multilevel/Phase Signal. vol. IT-28. IEEE Trans. Information Theory (1982)
7. Hae Geun Kim: Trellis-Coded M-ary Orthogonal Modulation. IEEE Symposium on Computer and Communications. pp.364-367. 1995

The Design and Implementation of Real-Time Environment Monitoring Systems Based on Wireless Sensor Networks

Kyung-Hoon Jung¹, Seok-Cheol Lee¹, Hyun-Suk Hwang², and Chang-Soo Kim^{1,*}

¹ PuKyong National University, Dept. of Computer Science, Korea
{jungkh, host2000}@mail1.pknu.ac.kr,
cskim@pknu.ac.kr

² PuKyong National University, Institute of Engineering Research, Korea
hhs@mail1.pknu.ac.kr

Abstract. This research focuses on the implementation of a real-time environment monitoring system for environment detection using wireless sensor networks. The purpose of our research is to construct the system on the real-time environment with the technology of environment monitoring systems and ubiquitous computing systems. Also, we present the monitoring system to provide a faster solution to prevent disasters through automatic machine controls in urgent situations. As the purpose of this study, we constructed simulation nodes with wireless sensor network devices and implemented a real-time monitoring system.

1 Introduction

The context-aware technology which is a core technology to construct a ubiquitous computing environment has recently become a growing interest. The representative technology of the context-aware is the Radio Frequency Identification and Ubiquitous Sensor Network (RFID/USN)[3][14][15]. In ubiquitous sensor network systems, the wireless node devices are installed on objects or places. The collected data through self-communication among them are transmitted to central nodes such as pan coordinators or sink nodes. Pan coordinators gather and analyze the data, send a control signal to control nodes such as actuators over threshold values, and adjust control conditions automatically. The USN system can be an automated network based on objects in all situations[8][9].

The USN has been utilized in robot control systems and monitoring systems, automatic temperature control systems and illuminate control systems in agricultural fields as well as other locations such as tracking system, smart homes, system, intrusion detection systems in applied applications. The research into the USN systems has proceeded through projects at universities in the U.S.A. For instances, the ecological system which has been developed by the Great Duck Island Project of U.C Berkeley installed sensor nodes at habitats of petrels and track their locations and moving paths[13][14]. Also, much research of monitoring system which can analyze crack states of buildings and bridges with sensor nodes has currently proceeded.

* Corresponding author.

Recently, USN systems can be applied to environmental problems and prevention of disasters there is overpopulation or in larger cities. In this paper, we will design and implement a real-time environment information system based on the USN to solve environmental problems that might cause large scale of disasters. Also, we will evaluate the performance of the presented system.

This paper is organized in the following manner. In the next section, we describe the basic technologies of ubiquitous computing and the real-time environment monitoring systems. In section 3, we present the hierarchy architecture of real-time environment control systems and explain their components. Also, we will experiment on the performance and the durability of sensor nodes in Section 4. Finally, we will summarize this research and will describe future work.

2 Related Research

2.1 Technologies of Ubiquitous Sensor Networks

The core technologies to implement ubiquitous sensor networks are classified into sensors, processors, communications, interfaces, and security[12][14]. First, sensors are the most important devices to sense variations of the environment as devices substituted by five senses of humans. The sensors devices are a core technology deriving objects-oriented computing environment from what has been human-oriented. Secondly, processors, which are devices corresponding to the human brain, process and analysis measured data by sensors. Processors in sensor networks can be implemented with only micro control units (MCU) executing essential functions. The low electricity-based MCU is an essential component to sustain durability of nodes. Next, communications and interfaces have interactions transmitting measured data to objects or humans through wire or wireless. Finally, ubiquitous computing systems have weaknesses in the information security. The security of USN has been researched in order to solve the weaknesses by using authentication and integrity of information[12][13].

2.2 Real-Time Environment Monitoring Systems

The goal of the environment monitoring systems is to minimize damages from disasters by monitoring and analyzing various environmental data on the real-time. However, in existing environmental information systems, humans who were expert in measuring gathered the data with analogy measurements at certain time intervals [8]. The data was dependant on the time it was measured, not data of fixed quantity with minimized errors in the real-time. For example, humans obtained data with measuring devices in the case of measuring exhaust gas from factories, and the collected data are time-dependant values and is not considering environmental variations. Therefore, environment monitoring systems which collect and analyze data on the real-time with sensor networks is required in the age of rapid variation.

3 Design and Implementation of Real-Time Environment Monitoring Systems

We constructed a real-time environment monitoring system based on the USN with the following four advantages. Our presented system has embedded operating systems

executing the simple tasks, which ensures self communications between nodes. It is implemented as low electricity, and has operations of stable nodes. Also, we implemented control and monitoring application module. Fig. 1 shows the hierarchical architecture of the real-time environment system.

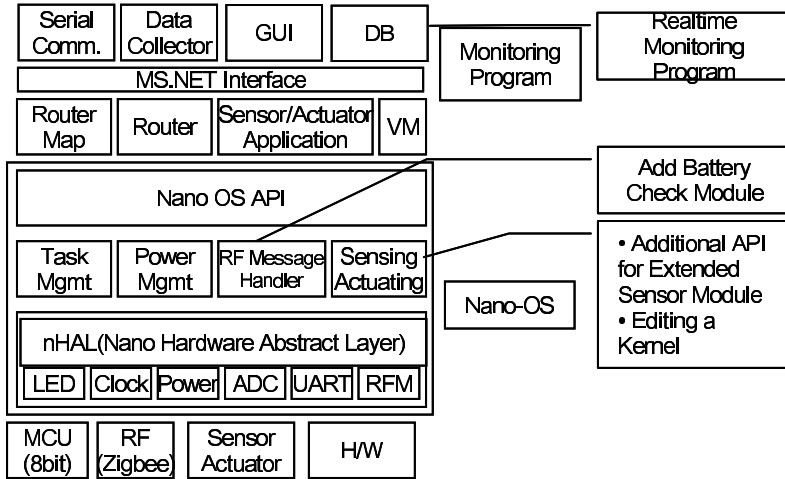


Fig. 1. The Hierarchical Architecture of the Real-time Environment System

3.1 Hardware Systems

We constructed the Nano-24 Development Kit Hardware System, which was developed by Electronics Telecommunications Research Institute in Korea (ETRI) and Octacomm Inc.[5][15]. The Nano-24 USN Kit is composed of the Main Board including RF module, the Sensor Board including gas, illuminate, temperature and

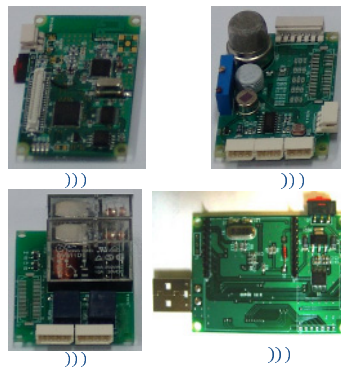


Fig. 2. Nano-24 Development Kit (a) Main Module (b) Sensor Module (c) Relay Module (d) Interface Module

humidity sensor, The Actuator Board can control machines by AC/DC relays, and the Interface Board which communicates with PC or monitoring systems using RS-232C standard interfaces. Figure 2 shows the four modules of the Nano-24 Development Kit.

3.2 Nano Real-Time Operating System with A/D Conversion Algorithm

Real-time embedded operating systems with simple task functions are used in USN systems[13][14]. We use a Nano-Q+ 1.5.1e version developed by ETRI. The Nano-Q+ is an operating system (OS) supporting hard real-time kernel and is being developed. In this paper, we reduce OS modules and add battery check modules to construct the effective monitoring system.

Analog to Digital (A/D) Conversion Algorithm. The A/D Conversion of sensors is executed by using registers related with ATmega128L made by the ATMel company. The procedures of the A/D conversion are follows:

- (1) The value of An ADMUX register and an ADCSRA register must be fixed. The ADMUX is a register variable for setting analog channels and standard voltages, and control ADC data registers to save results of A/D conversion. The ADCSRA is a register variable for setting pre-scalar, free running mode, and ADC-enabled.
- (2) A/D ending interrupt must be established for an interrupted type of A/D conversion. The ADIE bit of the ADCSRA variable is set up 1, and one bit of the SREG variable, which is an interrupt bit allowed, is set up 1.
- (3) The ADSC bit of ADCSRA variable is set up 1, and ADSC bit is set up 0 in the case of free running mode.
- (4) Wait for the interrupt occurrence of the A/D conversion ending
- (5) Read contents of the A/D data register

The results of ADC conversion have 0x0000~0x03FF(0~1023). The maximum value 0x03FF means ATmega128 has 10 bits of ADC address space.

```
SIGNAL(SIG_ADC){ //Get data from ADCL and ADCH
  adc_low_data=ADCL;
  adc_high_data=ADCH; }
void Init_ADC(void){ //Initialize ADC
  ADMUX=BM(REFS1)|BC(REFS0);
  ADMUX|=BM(MUX0);
  ADCSRA=BM(ADEN)|BM(ADSC)|BM(ADIE)|BM(ADPS0); }
```

Fig. 3. A/D Conversion Algorithm

3.3 Node Construction

We constructed a Pan Coordinator, one or more Sensor Nodes, and Actuator Nodes. Fig. 4 shows interactions between nodes.

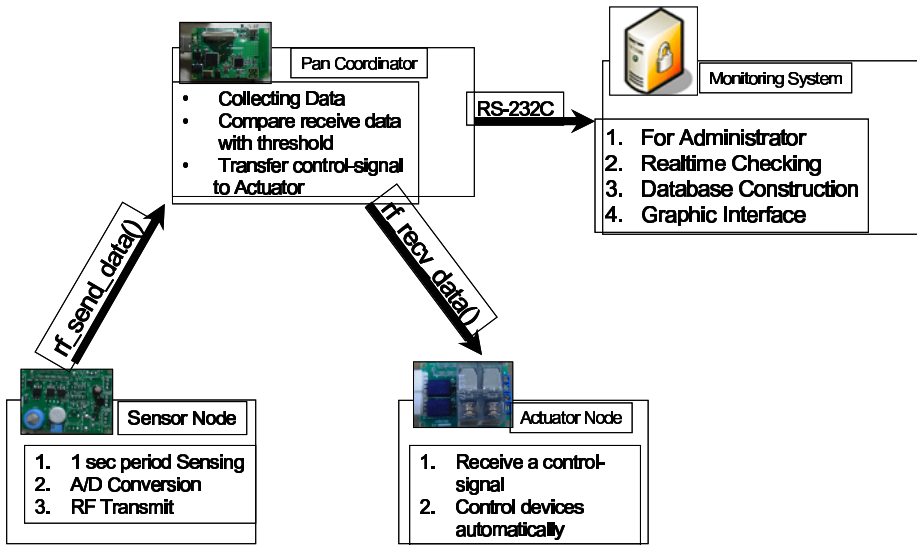


Fig. 4. The Construction of Nodes

```

/* initialize system values and then start multi-tasks. */
void *start(void *arg);
...
pthread_create(NULL,&attr,rf_net_scheduling(void *)0);
void *rf_net_scheduling(void *arg); /* network scheduling */
qplusn_tx_packet_queue_processing(); //Send Packet

/* receive task(rf-interrupt) */
void rf_recv_data (ADDRESS *srcAddr, INT8 nbyte, BYTE *data);
...
decode_indirect_packet(data,&route) // handling received message
/* send task */
void *rf_send_data (void *arg);
...
// Transfer a control-signal
if((int_data>250) && (MAIN_GAS_VALUE_STATUS==TURN_OFF))
    direct_actuator_op_cmd_transmit(1,GAS_VALUE,TURN_ON);
    MAIN_GAS_VALVE_STATUS = TURN_ON;85

```

Fig. 5. The Algorithm of the Pan Coordinator Nodes

Pan Coordinator Node. The pan coordinator node, which is located in the center of networks, collects data[4][14]. The transmitted data is compared with fixed threshold values and sends control signals to actuator nodes. Fig. 5 shows the algorithm for constructing the pan coordinator node.

Sensor Node. Sensor nodes are inputted into values extracted from the periodic sensing and transmit A/D conversion process and the data into pan coordinator nodes

by wireless[4]. In this paper, we constructed nodes with gas and illuminate sensor. Fig. 6 (left) shows the algorithm of sensor nodes.

Actuator Node. Actuator nodes receive control signals from pan coordinator nodes and set relays to ON/OFF[14]. Fig. 6 (right) shows the core algorithm of actuator nodes.

<pre>void *rf_net_scheduling(void *arg) GAS_SENSOR_POWER_ON(); /* sensor power on */ LIGHT_SENSOR_POWER_ON(); while(1) if (second_cnt==0) pthread_create(NULL, &attr, rf_send_data, (void *)0); //second_cnt= SENSOR_ADC_PERIOD; second_cnt= 10; /* switching */ pthread_ms_delay(1000); second_cnt--; void rf_send_pkt(void) int_data = get_gas_adc_raw_data(); sensor_gas = int_data; int_data = get_light_adc_raw_data(); sensor_light = int_data; nano_rf_send_pkt(&global_my_coordAddr, index+2, pBuffer, TX_OPT_ACK_REQ); halWait(50000);</pre>	<pre>void nIde_node_incomming_data_indication(ADDRESS *srcAddr, UINT8 nbyte, BYTE *pMsdU) ... if (packet_type == ACTUATOR_COMMAND_PACKET) actuator_type = (BYTE)(pMsdU[index]); actuator_op_mode = (BYTE)(pMsdU[index+1]); ... if (LIGHT_LAMP == actuator_type) /* LAMP Satatus Check */ actuator_operation(LIGHT_LAMP, actuator_op_mode); //rf_actuator_status_transmit(LIGHT_LAMP, actuator_op_mode); LED1_BLINKING(); else if (GAS_VALVE == actuator_type) /* LAMP Satatus Check */ if (MAIN_GAS_VALVE_STATUS != actuator_op_mode) MAIN_GAS_VALVE_STATUS = actuator_op_mode; actuator_operation(GAS_VALVE, actuator_op_mode);</pre>
---	---

Fig. 6. The Algorithm of Sensor Nodes (left) and Actuator Nodes (right)

3.4 Real-Time Monitoring Implementation

The real-time monitoring is a user interface program to express data extracted from RS232C standard serial communication interface based on the wireless sensor networks. The monitoring program consists of five modules which are serial communication module, data collection module, power management module, data storage module, and GUI module. The modules are developed by on the Microsoft Visual Studio .NET 2003. Fig. 7 shows the structure and the algorithm of the monitoring system.

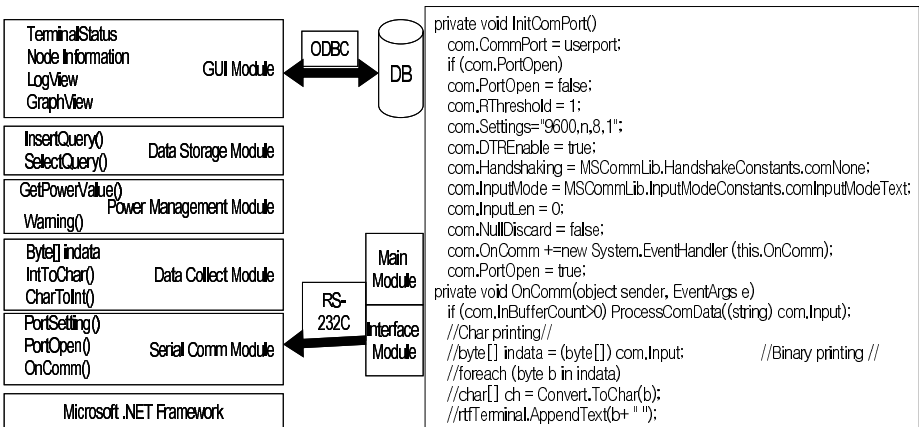


Fig. 7. The Monitoring System and Algorithm

Serial Communication Module. This module presents values extracted from RS-232C port. We implement the module by marshaling an 'AxMSCommLib.lib' file in the Visual Studio 6.0 tool because of without the library for controlling serial ports on the Microsoft .NET.

Data Collection Module. This module extracts data from sensors using characteristic function and expresses the collected values into following record structure.

SourceID	LightValue	GasValue	Voltage	Time	Reserved	Reserved	Padding
----------	------------	----------	---------	------	----------	----------	---------

Power Management Module. This module presents voltage values, and warning beeps if its values drop.

Data Storage Module. This module connects with the database systems to register sensor data and manage the related log data. The values extracted from the data collection module are executed with SQL query language.

GUI Module. This module is to support user friendly-forms.

4 Experiment Results

4.1 The Experiment of Threshold Values in Sensor Nodes

We defined the threshold values by extracting an upper and a lower value from sensor nodes. We experimented with an illuminate sensor A9060 and a Nap-55AE gas sensor to attain threshold values. The results which are attained from the outdoor and indoor with varied lighters as the defined time variation are showed in table 1.

Table 1. The A9060 Illuminate Sensor

Time	Outdoor	Indoor (fluorescent)	Indoor (fluorescent +lamp)
10:00 a.m.	772	703	820
15:00 p.m.	821	822	835
18:00 p.m.	355	790	822
21:00 p.m.	257	790	821

We experiment with lighter gas, butane gas, and LPG gas using a gas sensor NAP-55A. Table 2 shows values in normal and pouring gas as the types of gases.

Table 2. The NAP-55A E Gas Sensor

NAP-55A	Values in Normal	Values in Pouring Gas
Lighter Gas for a time	67	415
Butane Gas	66	633
LPG Gas in a Can	67	335

4.2 The Experiment of Sensor Performance

We experimented on an urgent imitation situation and measured time taken in controlling machine by relays using a timer module to attain more accurate results. We interrupt the illuminate sensor from light completely and attain the time taken until actuator node’s switch is in ON state. In the gas sensor, we pour into gas and describe the ON/OFF state of the motor and the ON state of the buzzer. The results are showed in Table 3. We knew our system can perceive the urgent values over threshold values.

Table 3. System Performance Results

Sensor	Illuminate Sensor	Gas Sensor
Machine State	Light On	Motor Off
Minimum Required Time	About 3.65 sec	About 4.7sec
Maximum Required Time	About 4.77 sec	About 8.5sec

4.3 The Experiment of Nodes Duality

We experimented on the sleep mode which provides the electric power for driving only basic MCU and RF circuit, and the active state mode which happens the periodic sensing to test the duality of nodes.

Fig. 8 shows results on the sleep mode (left) and on the active state mode (right). Nodes in the sleep mode are preserved to more than 15 hours, but are not preserved for more than 7 hours in the active stat mode. The result is analyzed because the power consumption quantity of the NAP-55A gas is about 30(mA) and the much power taken for deriving sensors needs.

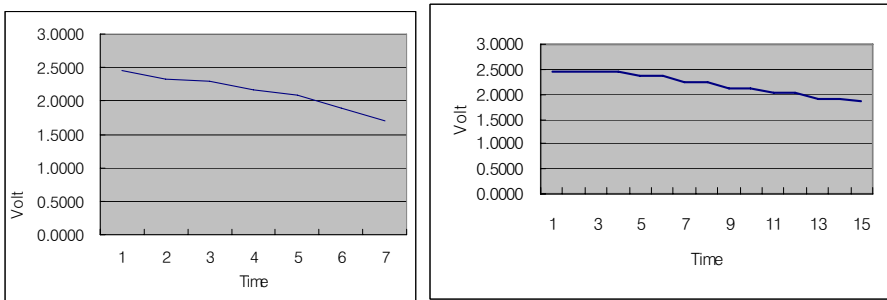


Fig. 8. Nodes Duality in Sleep Mode(left) and in Active State Mode(right)

5 Conclusion

We designed and implemented a real-time environment monitoring system based on the USN environment using wireless sensor networks. The system operates on the premise that pan coordinate node collects data extracted from sensor nodes, compares the

collected data with fixed threshold values, and control the machines automatically if the data is over the fixed threshold values. The advantages of our system include: (1) the independence of roles between nodes, (2) the automated system by self-communication function between nodes, and (3) the user-friendly monitoring system to access the environment information and to process the urgent situation more easily.

It needs some improvements in order to apply our system to a real field. First, there are some problems with the amount of electrical power consumption caused from driving sensors and causing low data integrity transmitted by wireless. Especially, this happens in the case of sensor networks that need many sensor nodes. Secondly, we know that it takes a long time when driving actuator nodes and controlling machines are using gas sensors. As a result, we need to develop an algorithm to lessen the time complexity. As future work, we are planning to apply the monitoring system to a real field and to continuously improve the performance of the system.

References

1. Akyildiz, Ian. F., Su, Weilian , Sankarasubramaniam, Yogesh, and Cayirci, Erdal: A Survey on Sensor Networks. *IEEE Communications Magazine*, August (2002) 103-114
2. Bulusu, et al.: Scalable Coordination for Wireless Sensor Networks: Self-Configuring Localization Systems. *ISCTA 2001*, Ambleside, U.K. (2001)
3. Choi, Y. H.: A plan of practical management in U-City. *Samsung SDS Inc* (2004)
4. Edgar, H. , Callaway, Jr.: *Wireless Sensor Networks Architectures and Protocols*. CRC Press (2004)
5. ETRI : Nano-Q+ Homepage (<http://www.qplus.or.kr>)
6. Kahn J. M., Katz R. H., and Pister K. S. J.: Next Century Challenges: Mobile Networking for Smart Dust. *Proc. ACM MobiCom '99*, Washington, DC (1999) . 271-78.
7. Kang, S.C.: The Future of a Sensor Network Stage. *Electronics and Information Center IT Report* (2003)
8. Kim, Dae young , Seong Ki Hong: A technology of smart sensor node operating system. the 97th TTA Journal 73-80
9. Korea Ministry of Information and Communication: The Basic Plan of Construction u-Sensor Network. Public data in Korea Ministry of Information and Communication (2004)
10. Korea Ministry of Information and Communication: Korea Ministry of Information and Communication Homepage(<http://www.mic.go.kr>)
11. Lee, Jae Yong: Ubiquitous Sensor Networking Technology. the 95th TTA Journal 78-83
12. Lee, Keun Ho : U-City Device Network , A strategy of U-City Construction and service model seminar, (2004)
13. Mohammad, Ilyas , Imad Mahgoub : *Handbook of Sensor Networks Compact Wireless and Wired Sensing Systems* , CRC Press , (2004)
14. Octacomm Inc. : The understanding of Embedded System – Development of Sensor Network. Octacomm Inc. , (2005)
15. Octacomm Inc. : OctaComm Homepage (<http://www.octacomm.net>)
16. Pyo, Cheol Sik : U-Sensor Network , ETRI , (2004).
17. Park, Seong Soo : An application development of Sensor network using Nano-24, Octacomm Inc. , (2004)

18. Pottie, G. J. and Kaiser, W. J.: Wireless Integrated Network Sensors, *Commun. ACM*, vol. 43, no. 5, May (2000) 551-58.
19. Son, Dae Rak : An application of field and specification of sensor, Hannam University , (2004)
20. Shen, C., Srisathapornpha, C. t, and Jaikaeo, C.: Sensor Information Networking Architecture and Applications, *IEEE Pers. Commun.*, (2001)52-59
21. Sohrabi, K. et al.: Protocols for Self-Organization of a Wireless Sensor Network, *IEEE Pers. Commun.*, (2000), 16-27.
22. Sinha, A. and Chandrakasan A.: Dynamic Power Management in Wireless Sensor Networks, *IEEE Design Test Comp.* (2001).
23. Woo, A., and Culler, D.: A Transmission Control Scheme for Media Access in Sensor Networks, *Proc. ACM MobiCom '01*, Rome, Italy, July (2001)221-35.

Ontology-Based Information Search in the Real World Using Web Services

Hyun-Suk Hwang¹, Kyoo-Seok Park², and Chang-Soo Kim^{3,*,**}

¹ PuKyong National University, Institute of Engineering Research, Korea
hhs@mail11.pknu.ac.kr

² KyungNam University, Div. of Computer Engineering, Korea
kspark@kyungnam.ac.kr

³ PuKyong National University, Dept. of Computer Science, Korea
cskim@pknu.ac.kr

Abstract. The ontology is an essential component to demonstrate the semantic Web, which is being described as the next Web generation. A semantic information search based on the ontology can provide the inferred and associated information between data. To develop an ontology application in the real world, we design the architecture of search systems based on the ontology using Web services. Our system consists of ontology modules, search procedure modules for searching, RDQL generator modules, and client modules for user interfaces. Also, we construct a hotel ontology integrated with the related terms and implement a search example with the defined ontology.

1 Introduction

The Semantic Web is a technology which adds well-defined documents on the Web for computers as well as people to understand the meaning of the documents more easily, and to automate the works such as information searches, interpretation, and integration.

The ontologies, which are an essential component of the semantic Web, define the common words and concepts used to describe and represent an area of knowledge [6]. The construction of ontologies in some domains such as travel [13], education [22], and medical data [12] has developed to integrate different data structure on the Web and to provide semantic information.

Most current Web sites have major limitations in finding search results and presenting them. The keyword-based search is not efficient because it often results in too many or too few hits. Also, the provided information has many redundant and unrelated results, so it takes a long time to find the information that users want. As a result, such searching is time consuming and often frustrates the Web search users [8].

Ontologies are linked to each other on the Web, and the linked ontologies provide the various applications with shared terminologies and understanding [12], [23]. Therefore, searching for information on the semantic Web will provide the search results with less redundancy, integrated terms, and inferred knowledge.

* Corresponding author.

** This research has been funded by Kyungnam University Masan, Korea.

The most researches related to the Semantic Web have been focused on the standards of Web Ontology Language(OWL), the ontology constructions, its infrastructure based on the crawlers [4], [7], and agents [8], [9], [10], [11]. However, searches based on the ontology are actively not supported for users in the real world because of there being a shortage of standards of the ontologies between the same domains, content based on the ontologies, and connection information to the related sites. Therefore, in this paper we will present the search architecture based on the ontology using Web services. We will design the ontology of a realistic hotel search domain with a search scenario and implement the system. Also, we will present the potential benefits in searching based on the ontologies in the real world.

This paper is organized in the following manner. In the next section, we describe the semantic search examples with ontology and explain how Web services work. In section 3, we present the architecture for searching based on the ontology. We define a hotel search ontology and implement a search example in the defined ontology in section 4. Finally, we summarize this research and describe future work.

2 Related Research

2.1 The Semantic Search with Ontology

The ontology defines the common words and concepts used to describe and represent an area of knowledge [20]. Until recently, the ontology has been researched by the AI ontology community. This has begun to move for applying ontologies on the Web, especially in the area of search and retrieval of information repositories.

The Semantic Web [1] is a technology to add information on the Web, to enable computers as well as people to understand the meaning of the Web documents more easily, and to automate the works such as information search, interpretation, and integration. The Semantic Web is an extension of the current Web for the next Web generation started at the W3C in 1998, when it was working on the OWL, a standardized Ontology-Specification-Language for the semantic Web.

Especially, the semantic search is an application of the Semantic Web to search and is designed to improve traditional Web searching. The search method using the ontology is gathering strength as another new way of Web searching [10], [15], [21].

Passin [16] presented the advantages of the ontology compared to conventional databases in data structure points of view. The search based on the ontologies can also provide the relationships between resources and can exchange data with other applications.

Sugumaran and Storey [21] presented the efficiency and the approach method of semantic-based search compared to keyword search method. The method of keyword-based search is to find the occurrence of string patterns specified by the users in component attributes and descriptions. On the other hand, the semantic-based approach is to use a natural language interface for generating initial queries and to augment the searching with domain information. To support the semantic search, the ontology of the domain model is constructed to integrate different sets of terms.

Applications related to e-commerce, information retrieval, portals and Web communities based on the semantic Web and the ontologies have been actively researched in a few years. Especially, the noteworthy projects with respect to this

work are researches such as the OntoSeek [9] regarding the technical structure and environment in the USA, the OntoWeb [19] regarding the semantic portal in the E.U., the OntoBroker [5] regarding the general structure in the Germany, and the OntoKnowledge [20] regarding the frame work in the Germany.

In addition, there has been research to develop ontologies in applications on the Web. Clark et al. [3] insist on the importance of semantic Web in higher education. They said the effect of education on the Web depends on how the newly emerging semantic Web is explored, and the effects will be profound if the semantic Web becomes as ubiquitous as the Web today. Domingue et al. [8] developed an Alice, which is an ontology-based e-commerce project. This aims to support the dynamic query interface of online users by using five ontologies describing customers, products, typical shopping tasks, external context, and 'Alice' media. The combined ontology-based queries and dynamic queries will provide end users with the benefit of looking for relationships in large volumes of data.

The recent researchers [12], [15], [17] have used a Protege Tool to construct data structures and contents for supporting the semantic Web. The OWL [18] is widely accepted as the standard language for sharing semantic Web contents. The OWL plug-in [14], [15], [17] is a complex Protege plug-in with functions to load and save OWL files in various formats, to edit OWL ontologies with custom-tailored graphical widgets, and to provide access to reasoning based on description logic.

2.2 Semantic Web Services

The existing distribution systems have had disadvantages that they could not communicate with different protocol to one another. Web services can integrate the distributed computing environment using SOAP protocol with XML documents, not Resource Description Framework (RDF) documents. The Web services need the interface among Web service providers, brokers, and consumers. The providers publish the developed Web services with Universal Description Discovery and Integration (UDDI), and the consumers bind with Web Service Description Language (WSDL) and Simple Object Access Protocol (SOAP).

Passin et. al., [16] insist that the new version of SOAP makes it more practical to encode RDF data in a SOAP message if the current Web is oriented toward the semantic Web with semantic RDF contents and varied agents. Dameron et. al., [4] propose an architecture allowing the manipulation of ontologies using Web services. This enables users to implement such services like ontology Web services and their interfaces on the semantic Web. However, the functions rely on existing Web services technologies like SOAP and WSDL.

3 Architecture of an Ontology Based Search

3.1 System Architecture

Fig. 1 illustrates the architecture of semantic search using Web services. The systems consist of the content provider, semantic Web services, and search client.

The semantic Web services system includes the ontology server, Web services with remote search procedures, and Web server. The ontology server includes different

ontologies in varied domains and the ontologies can be exploited by different semantic Web applications. The remote search procedures support varied search functions for applications on client sides. The procedures need to connect the Jena API for querying the RDF contents and RDF Query Language (RDQL) Generator and RDQL Generator. The Web server needs to process the values of properties inputted by the client on the Web.

The content providers download the defined ontology from the ontology server and create RDF instances. The ontologies can be exploited by different semantic Web applications.

The search clients can attain required results by inputting the values for searching on the Web through Web server. Also, the client system can develop the applications by calling remote search procedures by Web services system on the mobile and personal system.

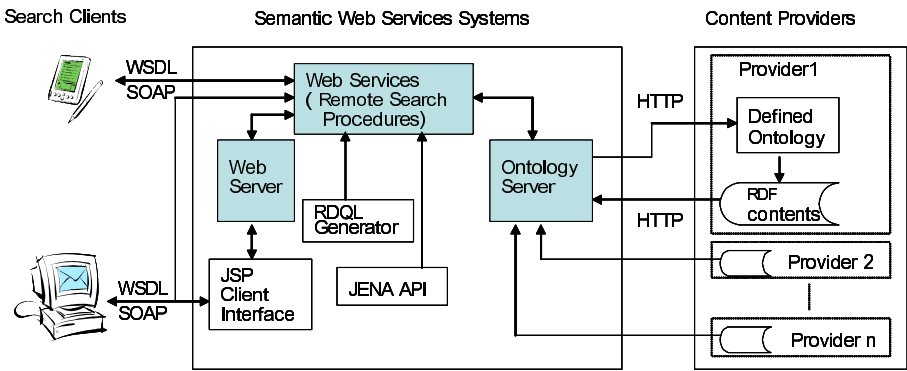


Fig. 1. Architecture of Ontology-Based Search

3.2 Implementation Module

Ontology Module. The ontologies could be defined by related industry. For example, Hotel Ontology could be provided by the hotel industry or hotel portal sites, and Geography Ontology could be defined by a government agency. The ontologies allow providers to get the defined ontology OWL files. The content providers submit the individual results as OWL or RDF files on their Web sites. The ontology server includes the interface form for providers to download the ontologies and to submit the URL with instances of the ontologies.

The ontologies can be constructed by using Protégé/OWL tool [14] which provides access to reasoning based on description logic and creates individuals with custom-tailored graphical widgets.

RDQL Generator Module. RDQL is one of the query languages for querying RDF contents. RDQL Generator generates the query string with the parameters of RDF models, properties needed by users for searching, and search conditions including subjects, properties, and objects which are elements of RDF statements.

Information Search Module. The ontology based information search can be divided into general keyword searching and ontology browsing searching. The general keyword method means the search based on data properties in total ontology structure, and the ontology browsing method means searches connected by object properties with relationships between classes. The browsing search allows users to search the information across the path connected between classes of the ontology.

The search procedures need the input information such as the URL of instances from providers, resulting properties from users, and RDQL query string. Also, the search procedures include data structures of properties of classes. Fig. 2 shows input values needed to attain search results.

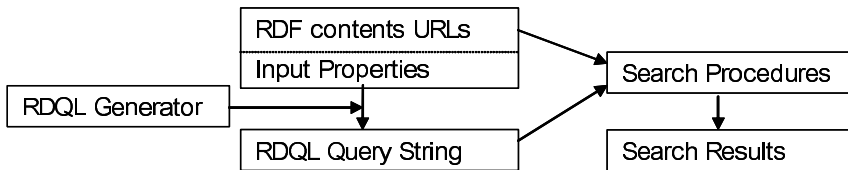


Fig. 2. Input Parameter of Search Procedures

Client Module. Clients can search for information based on the defined ontologies through their Web server and Web services. The clients need to submit input values for searches and then attain the search results. Also, clients can construct the applications on their client system by referencing WSDL files of Web services with defined ontologies. The applications can be generated on the mobile systems as well as personal systems.

4 Realizing a Hotel Search Ontology

We made up a search example based on the Hotel Ontology. We constructed the Web services with Java Web Services Development Pack 1.1 (JWS DP) including Tomcat server, and create the Hotel Ontology with the Protégé/OWL tool. We used the Jena API [2] to search requested information of users from the RDF-based contents generated by the Protégé/OWL. Also, we semantically searched RDF contents through RDQL, a Query Language for RDF. We used Java Server Page (JSP) to provide the user interface of information searches.

4.1 The Hotel Ontology

In this section, we show how we designed the ontology of a hotel domain from a search scenario which adjusts users' requirements. We imagined the search behavior of Web users who want to find a hotel with regards to some conditions like areas, room type, prices, and other facilities. The Web users try to find some candidate hotels of family suites with facilities like fitness centers and swimming pools, and they want to know them on for free. Therefore, we constructed the hotel ontology with more categorized classes including contract, service, room type, facility, and

rating. The ontology can provide more specific information by extracting information associated between data. For example, the users can obtain the information about what services are provided as the room type and whether the services are free or not. Fig. 3 shows the hierarchy of ontology for the hotel search and the relation connected by object properties between the classes in the defined ontology. The ‘Hotel’ Class is connected to the ‘Room’ class by the object property ‘hasRoom’, The ‘Room’ class is connected to the ‘Service’ class by the Property ‘hasService’.

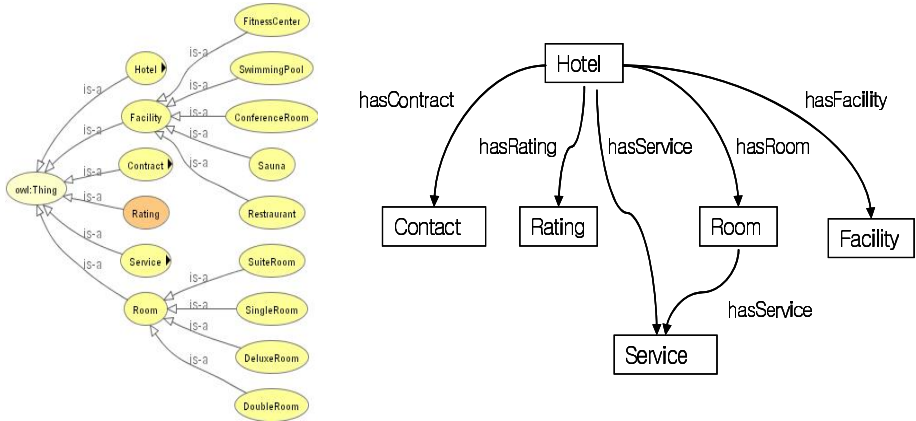


Fig. 3. The Fragment of the Hotel Ontology and the Associated Classes

4.2 Semantic Information Search

Our search systems support for clients to develop new search applications by accessing Web services. Fig. 4 shows the fragment of the WSDL needed to develop new applications and the part of calling search procedures of Web services on client systems.

<pre><?xml version="1.0" encoding="UTF-8" ?> <definitions xmlns="http://schemas.xmlsoap.org/wsdl/" xmlns:tns="http://localhost:8080/hotels/webservice/wsdl/webservice" xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:soap="http://schemas.xmlsoap.org/wsdl/soap/" name="webservice" targetNamespace="http://localhost:8080/hotels/webservice/wsdl/webservice"> <types> <schema xmlns="http://www.w3.org/2001/XMLSchema" xmlns:tns="http://localhost:8080/hotels/webservice/type/webservice" xmlns:xsd="http://www.w3.org/2001/XMLSchema" base="xsd:string" xmlns:soap="http://schemas.xmlsoap.org/wsdl/soap/" xmlns:encoding="http://schemas.xmlsoap.org/soap/encoding/" targetNamespace="http://localhost:8080/hotels/webservice/type/webservice"> <import namespace="http://schemas.xmlsoap.org/soap/encoding/" /> <complexType name="ArrayOfString"> <complexContent> <restriction base="soap-enc:Array"> <attribute ref="soap-enc:ArrayType" wsdl:arrayType="string" /> </restriction> </complexContent> </complexType> <complexType name="ArrayOfArrayOfString"> <complexContent> <restriction base="soap-enc:Array"> <attribute ref="soap-enc:ArrayType" wsdl:arrayType="ArrayOfString" /> </restriction> </complexContent> </complexType> </schema> </types> <message name="HotelsIF_hotelQuery"> <part name="String_1" type="xsd:string" /> <part name="ArrayOfString_2" type="tns:ArrayOfString" /> <part name="String_3" type="xsd:string" /> </message></pre>	<pre>try { hotels.Webservice_Impl webservice = new hotels.Webservice_Impl(); Stub stub = (Stub) webservice.getHotelsIFPort(); hotels.HotelsIF hotel = (hotels.HotelsIF) stub; for(int i=0; i<URL.length; i++) { disresults = hotel.hotelQuery(SURI[i], deResult, queryString); for(int k=0;k<disresults.length;k++) { if(disresults[k][0] != null) { for(int j=3;j<S;j++) { out.print(disresults[k][j]+ " "); } out.print("
"); } } } } catch (Exception ex) { ex.printStackTrace(); }</pre>
---	--

Fig. 4. WSDL and Calling a Search Procedure

Fig. 5(a) shows an example of a keyword search by users' requirement in the defined Hotel Ontology. The search can provide the abstract information from classes associated by hasFacility and hasService with search keywords like address, rating, and price. Fig. 5(b) shows the search example by ontology browsing, and the search provides the hierarchy of the terms and relations between classes to help the decision in finding candidate hotels. We can search some lists of hotels with some services followed by a room type through searching between connected classes.

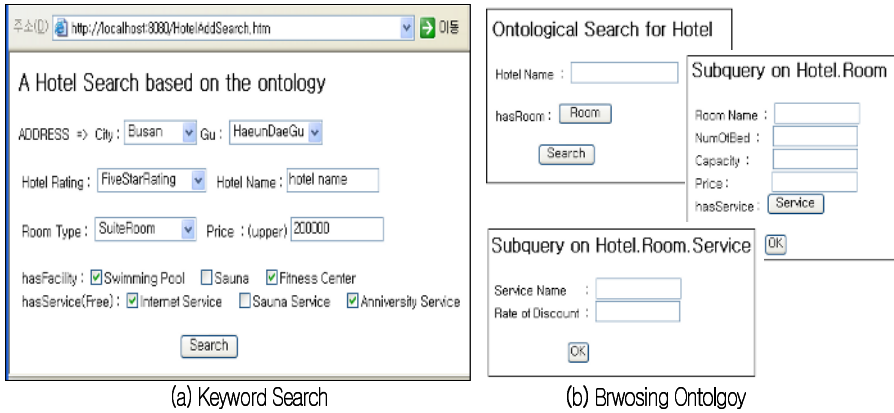


Fig. 5. Ontology-Based Search

4.3 Advantages of the Ontology-Based Search

The advantages of ontology-based search are follows. First, data integration on the Web can be accomplished by searching from ontology with integrated terms as domains, not extracting data from the different kind of database systems. Second, the ontology-based search provides more specific and hierarchical information by considering the relation between categorized classes, so finds the information from related data, not having been found from keyword search. In addition, it can do logic reasoning to discover unstated relationships in the data even though we do not implement the search with inference. Therefore, users can save the time and execute higher quality of Web search.

5 Conclusion

This paper presented the architecture of the information search based on the ontology using the Web services. The system consists of the ontology module, search procedures module, and client module. We constructed a standard ontology of hotel search domain with integration terms. Also, we implemented a hotel search example based on the defined ontology to improve the search of the current Web based on databases, which have problems such as the redundant and unrelated results and which are time consuming.

Our future works are as follows. The ontology of the hotel can be comprised in the variety viewpoint like travel theme categorized by mountain, beach, park, event and travel object like golf, business, and leisure with family and friends. Next, we can make up a portal site for the search based on the ontology to help the users to find some hotels adaptive to personal information.

Acknowledgement

This research was supported by the Program for the Training Graduate Students in Regional Innovation which was conducted by the Ministry of Commerce, Industry and Energy of the Korean Government.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American* Vol. 284, No. 5 (2001) 34-43
2. Carroll, J. Jeremy, Dickinson, Lan, Dollin Chris: Jena: Implementing the Semantic Web Recommendations. *Proceedings of the 13th International World Wide Web (2004)*
3. Clark, K., Parsia, B., Hendler, J.: Will the Semantic Web Change Education. *Journal of Interactive Media in Education (2004)*
4. Dameron, O., Natalya F., Knublauch, H., Musen, A.M.: Accessing and Manipulating Ontologies Using Web Services. *Third International Semantic Web Conference (ISWC2004), Hiroshima, Japan (2004)*
5. Decker, S., Erdmann, M., Fensel D., Studer R.: *Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information. Database Semantics: Semantic Issues in Multimedia Systems, Kluwer Academic Publisher (1999)*
6. Devedzic, Vladan: Understanding ontological engineering. *Communications of the ACM, Vol. 45, No. 4 (2002) 136-144*
7. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Sachs, J., Doshi, V., Reddivari, P., Peng, Y.: Swoogle: A Search and Metadata Engine for the Semantic Web. *Thirteenth ACM Conference on Information and Knowledge Management (CIKM'04), Washington DC, USA (2004)*
8. Domingue, J., Stutt, A., Martins, M.: Supporting Online Shopping through a Combination of Ontologies and Interface Metaphors. *International Journal of Human-Computer Studies 59 (2003)*
9. Guarino, Nicola, Masolo, Claudio, Vetere, Guido: *OntoSeek: Content-Based Access to the Web. IEEE Intelligent Systems, Vol. 14, No. 3 (1999) 70-80*
10. Guha, R., McCool, Rob, Miller, Eric: *Semantic Search. Proceedings of the 12th International Conference on World Wide Web (2003)*
11. Knublauch, Holger, Musen, Mark A, Rector, Alan, L.: *Editing Description Logic Ontologies with the Protégé OWL Plugin. International Workshop on Description Logics – DL (2004)*
12. Knublauch, Holger, Dameron, Olivier, Musen, Mark A.: *Weaving the Biomedical Semantic Web with the Protégé OWL Plugin. First International Workshop on Formal Biomedical Knowledge Representation, Whistler, BC, Canada (2004)*

13. Knublauch, Holger: Ontology-Driven Software Development in the Context of the Semantic Web: An Example Scenario with Protege/OWL. International Workshop on the Model-Driven Semantic Web, Monterey, CA (2004)
14. Knublauch, Holger, Protege OWL Plugin tutorial. 7th International Protege Conference, Bethesda, MD, <http://protege.stanford.edu/plugins/owl/documentation.html> (2004)
15. Knublauch, Holger, Ferguson, Ray W., Noy, Natalya F., Musen, Mark A.: The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. Third International Semantic Web Conference - ISWC (2004)
16. Passin, Thomas B., Explorer's Guide to the Semantic Web. Manning (2004)
17. Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., Wang, H., Wroe, C.: OWL Pizzas: Practical Experience of Teaching OWL-DL: Common Errors & Common Patterns. 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW) (2004)
18. Smith, Michael K., Welth Chris, McGuinness, Deborah L.: OWL Web Ontology Language Guide. [Http://www.w3.org/TR/owl-guide/](http://www.w3.org/TR/owl-guide/) (2003)
19. Spyns, Peter, Oberle, D., Volz, R., Zheng, J., Jarrar, M., Sure, Y., Studer R., Meersman, R.: OntoWeb: a Semantic Web Community Portal. In Proc. Fourth International Conference on Practical Aspects of Knowledge Management (PAKM), Vienna, Austria (2002)
20. Staab, Steffen, Studer, R., Schnurr, Hans-Peter, Sure, Y.: Knowledge Management : Knowledge Processes and Ontologies. IEEE Intelligent Systems Journal (2001)
21. Sugumaran, Vijayan, Story, V. C.: A Semantic-Based Approach to Component Retrieval. The DATABASE for Advances in Information Systems, Vol. 34, No. 3 (2003)
22. Tane, J., Schmitz, C., Stumme, G.: Semantic Resource Management for the Web: An E-Learning Application. Proceedings of the 13th international World Wide Web (2004)
23. Uschold, M. F., Jasper, R. j.: A Framework for Understanding and Classifying Ontology Applications. Proceedings of the IJCAI-99 workshops on Ontologies and Problem-Solving Mehtod (KRR5), Stockholm, Sweden (1999)

An Active Node Set Maintenance Scheme for Distributed Sensor Networks

Tae-Young Byun¹, Minsu Kim², Sungho Hwang², and Sung-Eok Jeon²

¹ School of Computer and Information Communications Engineering,
Catholic University of Daegu, Gyeongsan, Gyeongbuk, Korea
tybyun@cu.ac.kr

² School of Electrical and Computer Engineering,
The Georgia Institute of Technology, Atlanta, Georgia 30332-0250, USA
{mskim, sungho, sejeon}@ece.gatech.edu

Abstract. In this paper, we propose an energy-efficient coverage maintenance scheme for prolonging the lifetime of the sensor networks. Researchers are actively exploring advanced power conservation approaches for wireless sensor network and probabilistic approaches among them are preferred due to advantages of simplicity. However, these probabilistic approaches have limited usefulness because they can not ensure full area coverage. In the proposed scheme, each node computes the probability through the densities of its own coverage and keeps it adaptively to the report history. The performance of proposed scheme is investigated via computer simulations. Simulation results show that the proposed scheme is very simple nevertheless efficient to save the energy.

1 Introduction

Recently, the idea of wireless sensor networks has attracted a great deal of research attention due to wide-ranged potential applications that will be enabled by wireless sensor networks, such as battlefield surveillance, machine failure diagnosis, biological detection, home security, smart spaces, inventory tracking, and so on [7][9][10][13].

A wireless sensor network consists of tiny sensing devices, deployed in a region of interest. Each device has processing and wireless communication capabilities, which enable it to gather information from the environment and to generate and deliver report messages to the remote sink node. The sink node aggregates and analyzes the report message received and decides whether there is an unusual or concerned event occurrence in the deployed area. Considering the limited capabilities and vulnerable nature of an individual sensor, a wireless sensor network has a large number of sensors deployed in high density and thus redundancy can be exploited to increase data accuracy and system reliability. In a wireless sensor networks, energy source provided for sensors is usually battery power, which has not yet reached the stage for sensors to operate for a long time without recharging. Moreover, sensors are often intended to be deployed in remote or hostile environment, such as a battlefield or desert; it is undesirable or impossible to recharge or replace the battery power of all the sensors. However, long system lifetime is expected by many monitoring applications. The system lifetime, which is measured by the time until all nodes have

been drained out of their battery power or the network no longer provides an acceptable event detection ratio, directly affects network usefulness. Therefore, energy efficient design for extending system lifetime without sacrificing system reliability is one important challenge to the design of a large wireless sensor network. In wireless sensor networks, all nodes share common sensing tasks. This implies that not all sensors are required to perform the sensing task during the whole system lifetime. Turning off some nodes does not affect the overall system function as long as there are enough working nodes to assure it. Therefore, if we can schedule sensors to work alternatively, the system lifetime can be prolonged correspondingly; i.e. the system lifetime can be prolonged by exploiting redundancy.

A number of studies for reducing the power consumption of sensor network have been performed in recent years. These studies mainly focused on a data-aggregated routing algorithm [1 – 4] and energy efficient MAC protocols [6 – 8]. However, for more inherent solution to reduce energy consumption problem, the application level should be also considered [9][12]. In addition, the sensing area of each node may overlap because each link of a path to the sink node should be less than the radio radius. Therefore, it is important to reduce unnecessary traffic from overlapping sensing areas. Researchers are actively exploring probabilistic approaches due to its simplicity of distributed manner. However, these probabilistic approaches have limited usefulness because they can not ensure full area coverage. To guarantee the full coverage, our scheme maintains the report probability adaptively to the report history. Namely, each node calculates its report probability based only on the number of neighbors. Each node decides to report the message through the probability, and re-computes the probability using the report history. In other words, the probability of the node, which has transmitted at the previous report period, is decreased to avoid successive reporting. Reversely, uncovered area has more probability of sensing, that is covering, in the next period. The performance of proposed scheme is investigated via computer simulations. Simulation results show that our approaches reduce the report of redundant packets. Our paper is organized as follows. Section 2 reviews related works and section 3 introduces our scheme. In section 4 simulation results are presented, and finally, section 4 presents our conclusions.

2 Related Works

Sensor nodes are usually scattered in a sensor field. Each of these scattered sensor nodes is capable of collecting data and routing it back to the sink. Data are routed back to the sink by a multihop infrastructureless, and self-organized architecture. The sink may communicate with the task manager node via Internet or satellite.

In recent years, several important theoretical evaluations of coverage maintenance have been studied. These studies mainly analyze the distributed constructing and routing algorithms of a connected dominating set (CDS) [1 – 4]. In [4], Gao et al. present a randomized algorithm for maintaining a CDS with low overhead. The algorithm assumes the grid partition of the coverage and selects a small number of cluster heads. The work show that the total number selected has an approximation factor of $O(\sqrt{n})$ of the minimum theoretically possible. Wang et al. suggest a

geometric spanner algorithm that can be implemented in a distributed manner in [2]. The degree of node is limited by a positive constant, and the resulting backbone is a spanner for both hops and length. In [3], Alzoubi et al. describe a distributed algorithm for constructing a minimum connected dominating set (MCDS) with a constant approximation ratio of the minimum possible and linear time complexity. The above algorithms provide the theoretical limits and bounds of what is achievable with coverage maintenance. However, there is poor correlation between the spatial distance and reception rate, so assumptions based on geographic proximity between nodes do not necessarily hold in practice. Furthermore, the radio propagation is not circular, presenting non-isotropic properties. Therefore, the approximations under these assumptions may cause serious problems with algorithms that assume bidirectional connectivity [13].

The energy efficient protocols of MAC layer approach turn off the radios and do not transmit or receive of packets in a particular (usually small) timeframe. These protocols usually trade network delay for energy conservation because of the startup cost associated with turning the radios back on. Sohrabi and Pottie [6] propose a self-configuration and synchronization TDMA scheme at the single cluster. This work is more focused on the low-level synchronization necessary for network self-assembly, while we concentrate on efficient multihop topology formation. Sensor-MAC (S-MAC) [7] periodically turns off the radios of idle nodes and uses in-channel signaling to turn off radios that are not taking part in the current communication. More recent work [8] continues to explore MAC-level wake-up schemes. Most of the MAC schemes mentioned above can be applied to our scheme. Our scheme focuses on the method of generating report message, and thus independent to the MAC level approaches.

Another approach in reducing energy consumption has been to adaptively control the transmit power of the radio. The lazy scheduling proposed in Prabhakar et al. [9] transmits packets with the lowest possible transmit power for the longest possible time such that delay constraints are still met. Ramanathan and Rosales-Hain [10] proposed some distributed heuristics to adaptively adjust node transmit powers in response to topological changes caused by mobile nodes. This work assumes that a routing protocol is running at all times and provides basic neighbor information that is used to dynamically adjust transmit power. While power control can be very useful, particularly in asymmetric networks such as cellular telephony, their advantages are less pronounced in sensor networks [5]. In Xu et al. [11], GAF nodes use geographic location information to divide the network into fixed square grids. Nodes in each grid alternate between sleeping and listening, and there is always one node active to route packets per grid. Our scheme does not need any location aids since it is based on connectivity. Chen et al. [12] proposed SPAN, an energy efficient algorithm for topology maintenance, where nodes decide whether to sleep or join the backbone based on connectivity information supplied by a routing protocol. Our scheme does not depend on routing information nor need to modify the routing state; it decides whether to generate a report message or not based on adaptive report probability. In addition, our work does not presume a particular model of fairness or network capacity that the application requires.

3 Proposed Scheme

Due to the tight restrictions of the sensor node, low power consumption is one of the most important requirements. In addition, fairness is also a major requirement for construction of efficient sensor networks. To satisfy the requirements for self-organizing wireless sensor networks, we suggest a distributed scheme of controlling the transmission of sensing data by considering the geographical density of nodes. Namely, each node makes a report rule based on the geographical density and determines whether transmit the sensing data or not through the rule. In the proposed scheme, the rule is defined as a probabilistic approach. Therefore, each node determines whether transmit the sensing data or not using the report probability.

All nodes investigate the densities, share the densities with their neighbors, and compute the report probabilities. For the simplicity, we define the density of node i , denoted by $n_i.d$, as follows:

$$n_i.d = \frac{1}{\|n_i.nn\|} \quad \text{for } i = 1, 2, 3, \dots \quad (1)$$

where $n_i.nn$ is the set of neighbor IDs of i^{th} node.

In addition, the average density of neighbors can be expressed by

$$n_i.n\bar{d} = \frac{\sum_{k \in n_i.nn} n_k.d}{\|n_i.nn\|} \quad \text{for } i = 1, 2, 3, \dots \quad (2)$$

Using (1) and (2), the report probability of node i at first report period, denoted by $n_i.p_0$, can be defined as

$$n_i.p_0 = \begin{cases} 0 & : \|n_i.nn\| = 0 \\ \min(1, \alpha n_i.d + (1-\alpha)n_i.n\bar{d}) & : \|n_i.nn\| \neq 0 \end{cases} \quad \text{for } i = 1, 2, 3, \dots \quad (3)$$

where α is scaling factor and $\|n_i.nn\| = 0$ means that n_i has no neighbors. Note that the node without neighbors needs not to gather or transmit sensing data.

As mentioned above, probabilistic approaches may consume a long time for covering entire area due to their randomness. In our scheme, the report probability is maintained adaptively to the report history for solving these problems. The report history means whether a report is performed or not at previous report period, and can be expressed by

$$h(j) = \begin{cases} -1 & : \text{has transmitted at } (j-1)^{\text{th}} \text{ period} \\ 1 & : \text{has not transmitted at } (j-1)^{\text{th}} \text{ period} \end{cases} \quad (4)$$

Using (3) and (4), the report probability of node i at j^{th} report period, denoted by $n_i.p_j$, can be calculated as

$$n_i.p_j = n_i.p_{j-1} + h(j)\beta, \quad \text{for } i, j = 1, 2, 3, \dots \quad (5)$$

where β is the adaptivity factor.

To validate the performance of the proposed scheme, several terms are needed to be defined. The reachability is very important factor to determine the number of

nodes should be scattered over the sensor field. The reachability, denoted by RE , can be defined as the number of nodes that can deliver the sensing information to the sink and can be obtained by

$$RE = \frac{\eta}{N} \cdot 100 \quad (6)$$

where η denotes the number of nodes that have a path to the sink and N means the number of deployed nodes.

Sensing degree is another factor for sensor networks. We divide sensing field into 1×1 unit grids to compute sensing degree and coverage. Therefore, sensing degree can be computed by the average sensing redundancy of a grid. To explain the redundancy of a grid (x, y) , we define the redundancy function $f(i, x, y)$ as follows:

$$f(i, x, y) = \begin{cases} 1: \sqrt{(n_i \cdot x - (x + 0.5))^2 + (n_i \cdot y - (y + 0.5))^2} \leq r \\ 0: \sqrt{(n_i \cdot x - (x + 0.5))^2 + (n_i \cdot y - (y + 0.5))^2} > r \end{cases} \quad (7)$$

where r denotes the radio radius, and $n_i \cdot x$ and $n_i \cdot y$ are the x and y coordinates of i^{th} node, respectively. In addition, X and Y are the width and vertical length of sensor field, respectively.

Then, we can compute the density of a grid $d_{x,y}$ as follows:

$$d_{x,y} = \sum_{i=1}^N \sum_{x=1, y=1}^{X, Y} f(i, x, y) \quad (8)$$

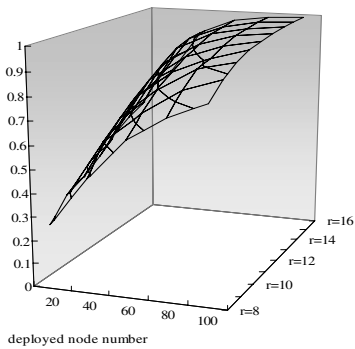
Next, sensing degree, denoted by SD , can be calculated by

$$SD = \frac{d_{x,y}}{XY} \quad (9)$$

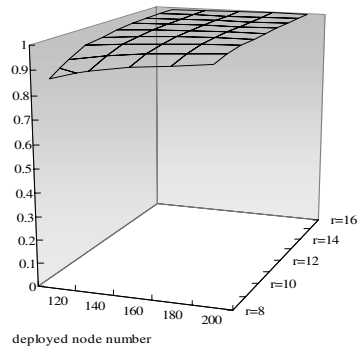
4 Simulations

To analyze the performance of our scheme, we carry out some experiments in static networks. We deploy 100 nodes in a square space (100 x 100). Nodes' x - and y -coordinates are set randomly. Each node has a sensing range of 15 meters and knows its neighbors. Note that the position, the neighbors are located, is not necessary in our scheme. We let each node decide whether to report or not based on its report probability. The decision of each node is visible to the neighbors. The nodes, which make decisions later, cannot "see" the nodes that have been turned off before. The current sensing coverage by active nodes is compared with the original one where all nodes are active. To calculate sensing coverage, we divide the space into $1\text{m} \times 1\text{m}$ unit grids. We assume an event occurs in each grid, with the event source located at the center of the grid. We investigate how many original nodes and how many active nodes can detect every event. In this experiment, we assume that the sensing coverage of node is similar to the radio coverage because the node can deliver the sensing information via only radio. In fact, the existence of an optimal transmission radius in the request-spreading process suggests an advantage in having a transmission radius larger than the sensing radius because the sensing radius directly affects the average

distance between area-dominant nodes. Moreover, enlarging the transmission radius can also benefit data-fusion schemes by allowing the construction of better-balanced trees. We plan to study sensor networks in which the sensing and transmission radius are different in the future. In our simulation, we assumed that the sink node is located in the center of the sensor field. Fig. 1 shows a 3D surface plot of the reachability in different sensing range and deployed node numbers. We change node density by varying the sensing range, denoted by r , from 6 to 16 and the deployed node number from 20 to 100 (Fig. 1a) and from 120 to 200 (Fig. 1b) in the same 100m×100m deployed area. From it, we can see that increasing the number of the deployed nodes and increasing the sensing range will result in more nodes being idle, which is consistent with our expectation. As shown in this Fig. 1b, the reachability reaches 1 when the number of deployed nodes is over 100. It indicates that most of deployed node can find route to the sink node when the number of deployed nodes is over 100.

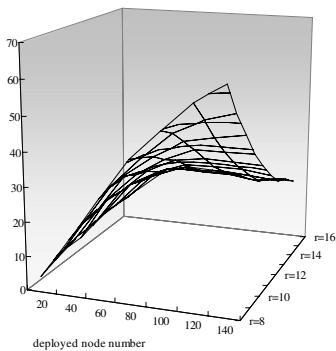


(a) Deployed node number: 20 ~ 100

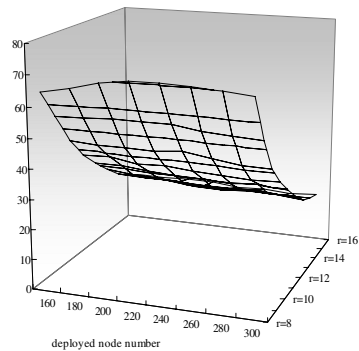


(b) Deployed node number: 120 ~ 200

Fig. 1. Reachability vs. node density



(a) Deployed node number: 20 ~ 140



(b) Deployed node number: 140 ~ 300

Fig. 2. Number of active nodes vs. node density

Fig. 2 shows 3D surface plot of the number of active nodes in different sensing range and deployed node numbers. We change node density by varying the sensing range, denoted by r , from 6 to 16 and the deployed node number from 20 to 100 (Fig. 2a) and from 120 to 200 (Fig. 2b) in the same 100×100 deployed area. We can also see that the active node number remain constant over different deployed node number when the sensing range and deployed area are fixed. These trends can be observed more precisely as illustrated in Fig. 3a. It means that the calculation of idle node is very easy using our scheme and can be easily applied to various MAC level power saving approaches.

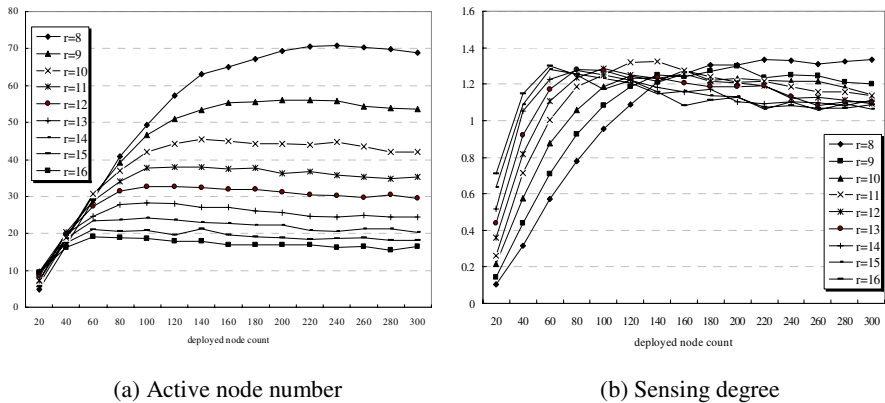


Fig. 3. Active node number and sensing degree vs. node density

We also investigate the sensing degree vs. node density. As shown in Fig. 3b, since nodes deployed on the sensing area densely enough, the sensing degree approximates 1 ~ 1.4. This indicates that our scheme reaches optimal solution. In Fig. 4, sensing degree of out scheme is compared with original sensing degree.

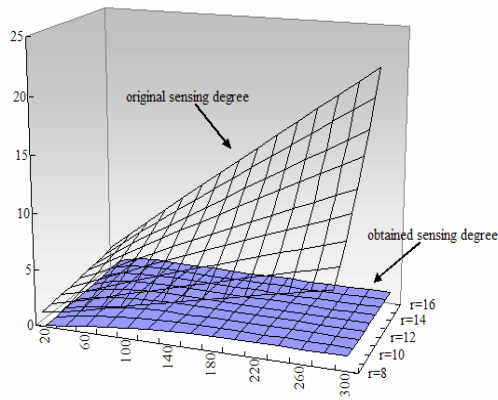


Fig. 4. Sensing degree reduction vs. node density

Fig. 5 presents the same effectiveness but from the different view: the ratio of the covered area. We still divide the space into 1×1 unit grids as mentioned earlier. An event occurs in each grid, with the event source located at the center of the grid. We

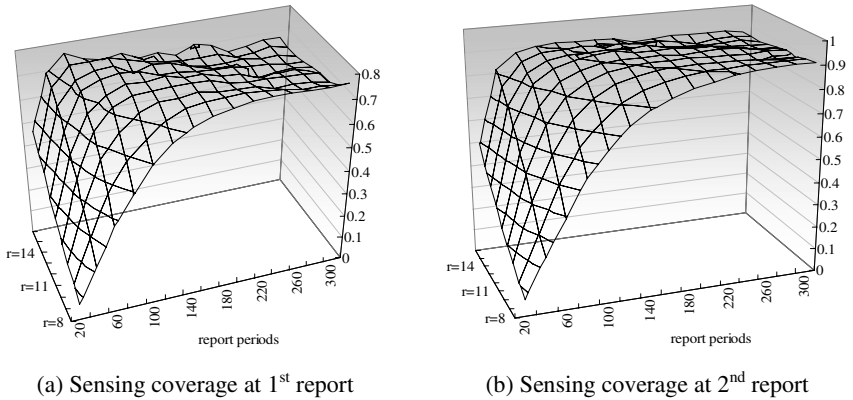
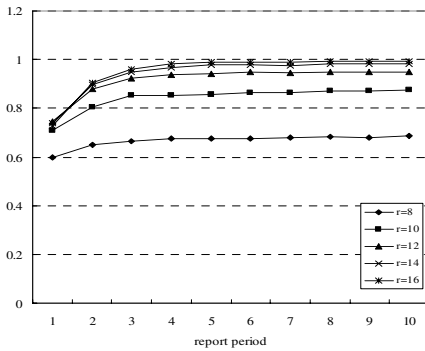
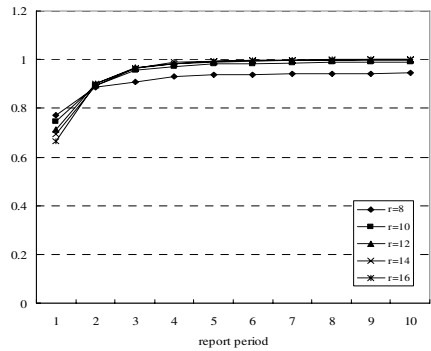


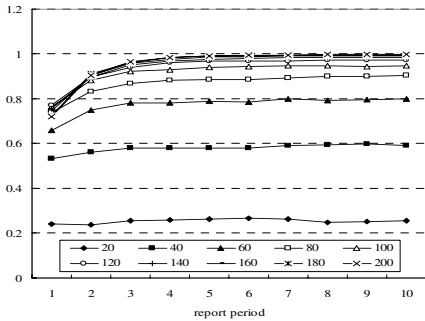
Fig. 5. Sensing coverage vs. node density



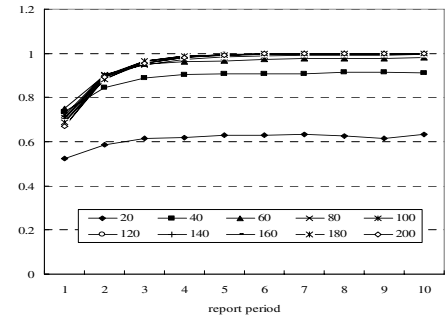
(a) Sensing coverage vs. range (N=100)



(b) Sensing coverage vs. range (N=200)



(c) Sensing coverage vs. node density ($r = 12$)



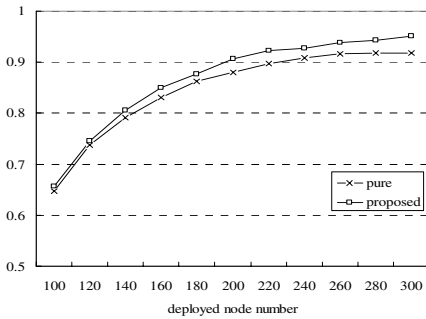
(d) Sensing coverage vs. node density ($r = 18$)

Fig. 6. Sensing coverage vs. report period

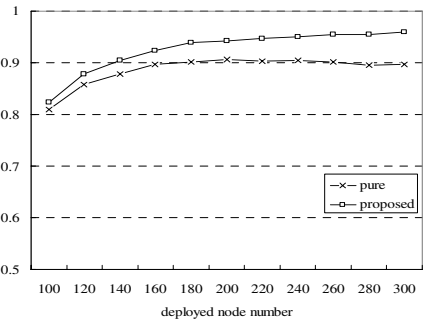
investigate the ratio of the grid number reached by active nodes to the total number of grids when sensing range is in from 8 to 16. As illustrated in the figures, most of the area, above 80%, can be covered by our scheme.

In the above figures, a pure probabilistic approaches and the proposed scheme don't have differences because the proposed scheme performed actively to the report history. Therefore, Fig. 6 and Fig. 7 show the sensing coverage vs. report history. First, Fig. 6 depicts the ratio of the covered area with changing the report period. We investigate the ratio of the grid number reached by active nodes to the total number of grids using pure probabilistic approaches when report period is in from 1 to 10. As illustrated in the figures, most of the area, after 4th period, can be covered by pure probabilistic approaches.

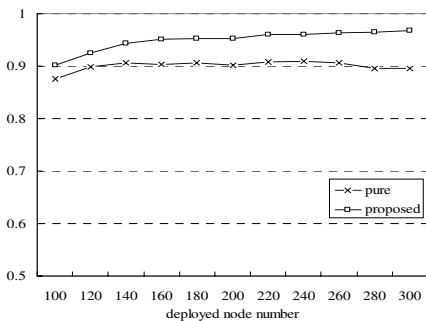
Next, we are going to compare the proposed scheme with pure probabilistic approach at the 2nd report period. Fig. 6 shows the ratio of the covered area of pure probabilistic approach and proposed scheme with changing the report period. As showed in these figures, proposed scheme reaches at 1 more quickly than pure probabilistic approach.



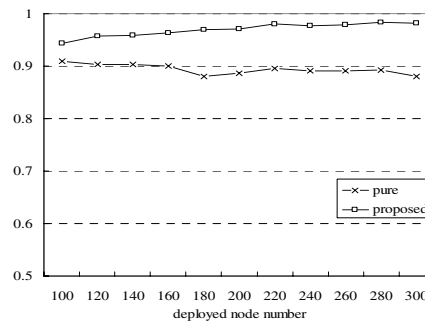
(a) Sensing coverage vs. node density (r= 8)



(b) Sensing coverage vs. node density (r = 10)



(c) Sensing coverage vs. node density (r= 12)



(d) Sensing coverage vs. node density (r = 14)

Fig. 7. Sensing coverage at 2nd report period

5 Conclusions

This paper presents a scheme for reducing the redundant power consumption in self-organizing wireless sensor networks. Our scheme computes adaptive report probability and controls a packet report through the probability. The performance of our scheme is investigated deeply via computer simulations, and the results show that our scheme is very simple nevertheless efficient to save energy.

References

- [1] J. Gao, L.J. Guibas, J. Hershburger, L. Zhang, and A. Zhu, "Geometric Spanner for Routing in Mobile Networks," Proc. Second ACM Symp. Mobile Ad Hoc Networking and Computing (MobiHoc 01), pp. 45-55, Oct. 2001.
- [2] J. Gao, L.J. Guibas, J. Hershburger, L. Zhang, and A. Zhu, "Discrete and Computational Geometry," Proc. Second ACM Symp. Mobile Ad Hoc Networking and Computing (MobiHoc 01), vol. 30, no. 1, pp. 45-65, 2003.
- [3] K.M. Alzoubi, P.-J. Wan, and O. Frieder, "Message-Optimal Connected-Dominating-Set Construction for Routing in Mobile Ad Hoc Networks," Proc. Third ACM Int'l Symp. Mobile Ad Hoc Networking and Computing (MobiHoc), June 2002.
- [4] Y. Wang and X.-Y. Li, "Geometric Spanners for Wireless Ad Hoc Networks," Proc. 22nd Int'l Conf. Distributed Computing Systems (ICDCS 2002), July 2002.
- [5] A. Cerpa, N. Busek, and D. Estrin, "SCALE: A Tool for Simple Connectivity Assessment in Lossy Environments," Technical Report CENS Technical Report 0021, Center for Embedded Networked Sensing, Univ. of California, Los Angeles, Sept. 2003.
- [6] K. Sohrabi and G. Pottie, "Performance of a Novel Self-Organization Protocol for Wireless Ad Hoc Sensor Networks," Proc. IEEE Vehicular Technology Conf., Sept. 2000.
- [7] W. Ye, J. Heidemann, and D. Estrin, "An Energy-Efficient MAC Protocol for Wireless Sensor Networks," Proc. 21st Ann. Joint Conf. IEEE Computer and Comm. Soc. (INFOCOM), pp. 1567-1576, June 2002.
- [8] R. Zheng, J.C. Hou, and L. Sha, "Asynchronous Wakeup for Ad Hoc Networks," ACM Int'l Symp. Mobile Ad Hoc Networking and Computing, June 2003.
- [9] B. Prabhakar, E. Uysal-Biyikoglu, and A.E. Gamal, "Energy-Efficient Transmission over a Wireless Link Via Lazy Packet Scheduling," Proc. 20th Ann. Joint Conf. IEEE Computer and Comm. Soc. (INFOCOM), pp. 386-394, Apr. 2001.
- [10] R. Ramanathan and R. Rosales-Hain, "Topology Control of Multihop Wireless Networks Using Transmit Power Adjustment," Proc. 19th Ann. Joint Conf. IEEE Computer and Comm. Soc. (INFOCOM), pp. 404-413, Mar. 2000.
- [11] Y. Xu, J. Heidemann, and D. Estrin, "Geography-Informed Energy Conservation for Ad Hoc Routing," Proc. Seventh Ann. ACM/IEEE Int'l Conf. Mobile Computing and Networking (MobiCom), pp. 70-84, July 2001.
- [12] B. Chen, K. Jamieson, H. Balakrishnan, and R. Morris, "Span: An Energy-Efficient Coordination Algorithm for Topology Maintenance in Ad Hoc Wireless Networks," Proc. Seventh Ann. ACM/IEEE Int'l Conf. Mobile Computing and Networking (MobiCom), pp. 85-96, July 2001.
- [13] A. Cerpa and D. Estrin, "ASCENT: Adaptive Self-Configuring sEnSOr Networks Topologies," IEEE Transaction on Mobile Computing and Networking, vol. 3, issue. 3, pp. 272-285, July 2004.

Intelligent Information Search Mechanism Using Filtering and NFC Based on Multi-agents in the Distributed Environment

Subong Yi, Bobby D. Gerardo, Young-Seok Lee, and Jaewan Lee

School of Electronic and Information Engineering, Kunsan National University,
68 Miryong-dong, Kunsan, Chonbuk 573-701, South Korea
{subongyi, bgerardo, leeys, jwlee}@kunsan.ac.kr

Abstract. Intelligent search agent is popularly used for searching relevant information in the Internet and there are lots of tools that are used to satisfy the needs of the users. Since there is no sufficient cooperation among the agents and they are independent to each other, it is difficult to make an efficient search of information in the distributed environment. Therefore, a typical search agent is difficult to use and can contain irrelevant information for the users. To solve these problems, we use the CORBA architecture to create an agency in the broker agent and provide more reliable information to the users. Also, the proposed intelligent information search system use the NFC and filtering techniques through the multi-agents for fast and reliable search of information.

1 Introduction

Nowadays, we are already experiencing the high-speed networks which are the medium for the Internet and other distributed information systems. Accompanied by the existence of the Internet, lot of information has increased and the work of the users who search for relevant information also increased rapidly. There is much information we can gather on the Internet and some information are not relevant to the user's request, which makes it difficult for the user to search necessary information. It is common that users want to search for information on a fast and efficient way while avoiding the irrelevant information to be processed that can be time-consuming. Many researches proposed efficient search of information. Some developed an intelligent agent for searching of information on the Internet and introduced cooperative multi-agents for searching of information. A push and pull method of information provides an efficient utilization of network resources [4].

Moreover, existing researches use typical agent architecture for searching information from information resources that operate independently and are dependent to each platform [1][2]. The communication and cooperation of agent are inefficient. Consequently, the reliability of information that is gathered in the distributed environment becomes low. Many researches study neural network and it is widely used in research task. One example is an agent which uses neural network that improves the accuracy and reliability of information, but has a longer processing time [5].

In order to solve these problems, we proposed an information search multi-agent system. The proposed system use CORBA to provide relevant information to users

and transparency of the system. The broker agent creates agency which communicates with the multi-agents to provide information. Internet Inter-ORB Protocol (IIOP) is used for the communications of multi-agent to send multiple packages of messages. In order to acquire accurate information, neural network and fuzzy technique are used. In addition, it uses filtering of the clustered data to reduce processing time. An agent which does the push service and manages the large resources in the distributed environment is also presented. The goals of the agent are to have an efficient resource management and provide fast processing of information.

2 Related Works

Information search in the Internet is a popular topic of research which includes reliability and efficiency. We gathered some related issues about neural networks and the multi-agents and discuss these in the following subsections.

2.1 Neural Networks

A neural network is an interconnected group of artificial or biological neurons. It is possible to differentiate between two major groups of neural networks. It is applicable in every situation in which a relationship between the predictor variables or inputs and predicted variables or outputs exists, even when that relationship is very complex and not easy to articulate in the usual terms of correlations. A few representative examples of problems to which neural network analysis has been applied successfully are in finance, medicine, engineering, geology and physics. There are a lot of researches which uses the neural network in information search. Two different approaches combining fuzzy genetic algorithms and the preprocessing stage of classification called feature selection is used for the information retrieval [1] where the processed document is rank by relevance of the information. A fuzzy clustering based on mobile agent is used for information search [5]. Here, a search mechanism based on ontology using a mobile agent to increase the accuracy rate of information is used. A cooperation of multi-agent is used to provide a reliable search mechanism in the distributed system [6].

2.2 Multi-agent

A multi-agent system (MAS) is a system composed of several agents, capable of mutual interaction. The agents are considered to be autonomous entities such as software programs or robots. Their interactions can be either cooperative or selfish. That is, the agents can share a common goal or they can pursue their own interests. An intelligent information retrieval (IIR) agent is one of a solution to the information overloading [2]. The multi-agent system identifies the desirable features of an IIR agent, including intelligent search, navigation guide, auto-notification, personal information management, personal preferred interface, and tools for easy page-reading. An information filtering approaches that are distributed with respect to knowledge or functionality, to overcome the limitations of single-agent centralized information filtering [3] is used. Large-scale experimental studies like in the Text Retrieval

Conference (TREC) are also presented to illustrate the advantages of distributed filtering as well as to compare the different distributed approaches.

3 Intelligent Information Search Agent (IISA) System

We proposed an Intelligent Information Search Agent (IISA) System, which is an information search system that produces more accurate information and process the information intelligently. Figure 1 presents the architecture of the proposed agent system which consists of three classes. These classes are interface agent, broker agent and resource agent.

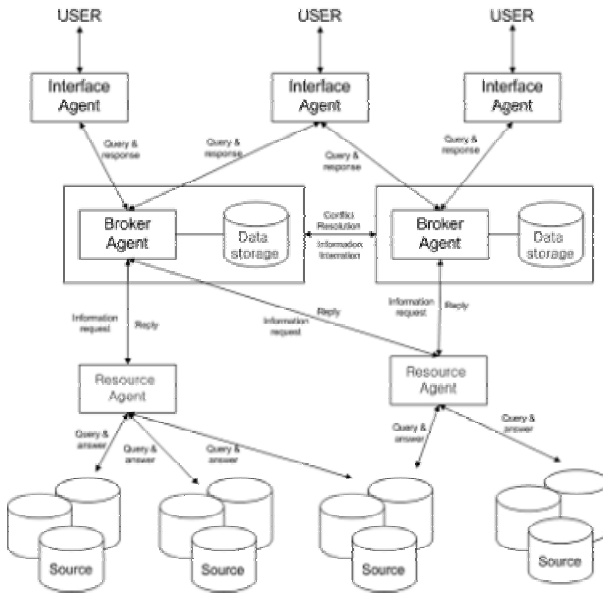


Fig. 1. Architecture of Intelligent Information Search Agent System

The interface agent which is a primary class takes charge of the input of the user for request of search. The resource agent manages the information resource contents. It collects information which is necessary and then sends it to broker agent for extracting the information. The broker agent acts as medium of communication between the interface and resource agent which manages the task of other agents and communicates by query language. In order to respond to the request of the user, the interface agent manages the request for information. The interaction of resource agent and broker agent leads to providing relevant information to the user.

3.1 Interface Agent of IISA

The interface agent interacts with the user and the IISA system. The inputs for the interface agent are the request of the user for information and then it stores the log

information of the user. After the interface agent process the request of the user then it shows the result of the process as output. The user's requests, interests and log information are delivered to broker agent and it processes the query of the user. The proposed system includes a resource agent that uses a data mining technique to gather the relevant data. The information which gets from the broker agent will be presented in a message format or graphical form to become easy to understand by the user.

3.2 Resource Agent of IISA

Resource agent manages the sources of information in the distributed environment. In order to gather the relevant and updated information, it uses the user log information of the interface agent. This log information is delivered to resource agent and it searches updated information about the previous log information. Then, the resource agent pushes the updated information to the broker agent. This is stored by the broker agent in the management table. Also, the resource agent does the resource monitoring in the dynamic distributed environment.

3.3 Broker Agent of IISA

Broker agent solves the heterogeneity and improves the transparency of the distributed environment by using CORBA architecture. One of the important goals of broker agent is to interact with the multi-agent to provide more reliable information. The broker agent creates an agency in order to communicate with the multi-agent agency (MAA). Figure 2 presents the broker agent architecture. After creating the agency, it performs the data mining using Fuzzy Neural Clustering (FNC). Figure 3 presents the process of extracting the information by using the FNC and filtering. In the first process, the fuzzy clustering is performed. The next process is filtering the clustered data by selecting the more interesting data with the user's request. This eliminates irrelevant data and improves the efficiency to extract information. The last process is using the neural network on the filtered data. This uses the self-organizing maps (SOM) to learn from data and then provide information. SOM is a good method to learn from data and to provide information but time processing is longer so we use filtering process to make it more efficient by not adding the irrelevant data in SOM. This approach reduces the processing time.

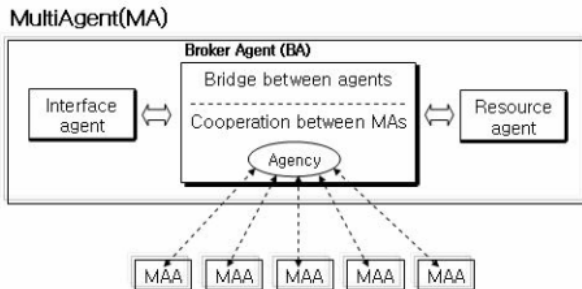


Fig. 2. Architecture of Broker Agent

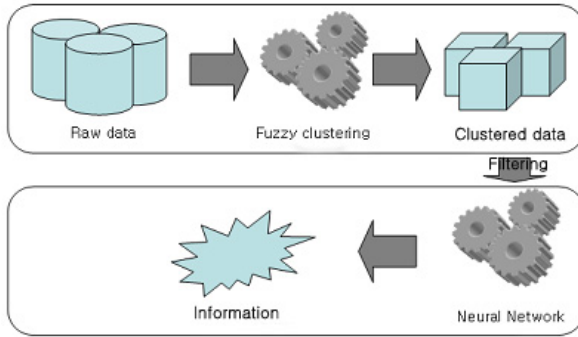


Fig. 3. NFC and filter processing of data

Agency. The agency is used for interaction of the multi-agent to provide data and mining process to produce information. The multi-agent interacts with the agency and it uses NFC and filtering processes. All information is stored in a stack which is the management table. This information will be accessed by the interface agent in case the user request for information.

FNC and Filtering. We present a technique using FNC and filtering of data. Our proposed algorithm enhances the speed of processing compared with the other existing information search technique that uses neural network. Figure 4 illustrates the architecture of the FNC and filtering. First, we use the Fuzzy C-Means (FCM) which is the second layer of the architecture of FNC. After the FCM, we filter the clustered data which occurred between the second layer and third layer. SOM is presented from the third layer until the fifth layer which is used to learn from the data and process it into information. In Equation 1 until Equation 5, we present the formula for fuzzy clustering. Equation 6 presents the formula of filtering the clustered data to be processed in SOM.

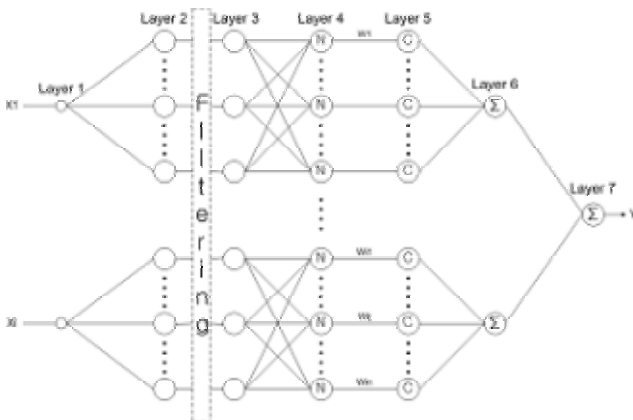


Fig. 4. Architecture of FNC

$$J(u_{ik}, v_i) = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m (d_{ik})^2 \quad (1)$$

$$d_{ik} = d(x_k - v_i) = \left[\sum_{j=1}^L (x_{kj} - v_{ij})^2 \right]^{1/2} \quad (2)$$

$$v_i = \{v_{i1}, v_{i2}, \dots, v_{ij}, \dots, v_{iL}\} \quad (3)$$

$$v_{ij} = \frac{\sum_{k=1}^n (u_{ik})^m x_{kj}}{\sum_{k=1}^n (u_{ik})^m} \quad (4)$$

$$u_{ik} = \frac{(1/\|x_k - v_i\|^2)^{1/m-1}}{\sum_{j=1}^c (1/\|x_k - v_j\|^2)^{1/m-1}} = \frac{1}{\sum_{j=1}^c \left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|} \right)^{2/m-1}} = \frac{1}{\sum_{j=1}^c \left(\frac{d_{ik}}{d_{jk}} \right)^{2/m-1}}$$

In this point, $d_{ik} = d(x_k - v_i) = \left[\sum_{j=1}^L (x_{kj} - v_{ij})^2 \right]^{1/2}$ (5)

$$DF_i = |V_i - Ix_i| \leq F_w (0.2 \sim 0.8) \quad (6)$$

```

set a number of cluster
c(2<=c<n), m(1<m<∞)
Initialize Make U-matrix

$$u_{ik}^{(r+1)} = \frac{1}{\sum_{j=1}^c \left( \frac{d_{ik}^{(r)}}{d_{jk}^{(r)}} \right)^{2/m-1}}$$
 for  $I_k=0$ 
Calculate the center v
Calculate distance data and cluster
//Update the partition matrix U
Goes through all input sample and calculates membership function
according to the cluster centers.
Calculates the fuzzy objective function values given center and U
Filtering
 $F_i = V_i - Ix_i \leq F_w (0.2 \sim 0.8)$ 
input a result:
Initialize all weights and biases in network;
while terminating condition is not satisfied {
  for each training sample X in samples { //Propagate the inputs forward for
  SOM unit j {
     $I_j = \sum_i w_{ij} O_i + O_j$ ;
    // compute the net input of unit j with respect to the previous layer, I
     $O_j = \frac{1}{1+e^{-I_j}}$ ; // compute the output of each unit j
    // for each unit j in the hidden layers, from the last to the first layer
    Errj =  $O_j(1-O_j) \sum_k Err_k w_{kj}$ ;
    // compute the error with respect to the next higher layer, k
    for each weight  $w_{ij}$  in network {
       $\Delta w_{ij} = (O_j - I_j) w_{ij}$ ; //weight increment
       $w_{ij}^{t+1} = w_{ij}^t + \Delta w_{ij}$ ; // weight update } }
  }
  While terminating condition is not satisfied {
    similarity set of A and B():
    calculate distance from near set():
    calculate errors(): }
  // compute the similarity using domain
  function similarity set of A and B() {
    if  $y = \text{Min}(U_i(x), U_j(x))$ 
      simij =  $U_{i,s}(x,y)$ 
    else }
  function calculate distance from near set() {
    S =  $\text{Max}(\text{Value}_{A_1}, \text{Value}_{B_1}, \dots, \text{Value}_{A_n})$  }
  function calculate errors() {
    err =  $\text{Value}_{s,1} \times (A_s \cap A_1 / A_s)$  }

```

Fig. 5. Algorithm for FNC and filtering

Procedures for FNC and Filtering:

1. Input value of the user by requesting the information
2. Fuzzy clustering using FCM
3. Filtering of the clustered data
4. Result of third procedure is gathered to be processed in SOM for learning of data
5. Calculates the weight value in the hidden layer of the SOM
6. Repeat the fifth procedure in the next hidden layer
7. Update the weight value from the result of the sixth procedure
8. Repeat fifth to seventh procedure until the weight value is more than the critical value. If the weight value is lesser than the critical value then end the procedure.

4 Performance Evaluation

The simulation environment of the proposed system consists of multi-agent and it has object modules for agency. These multi-agents provide information processing through the interaction to each other. To perform the heterogeneity of the system, we use the Solaris 9 and Windows XP operating systems. The information processing is done in the heterogeneous environment by using different operating systems and uses the Visibroker software by Borland which follows the OMG CORBA standards. The programming languages used in the development of the multi-agents are Java and C++. Figure 6 presents the system simulation environment.

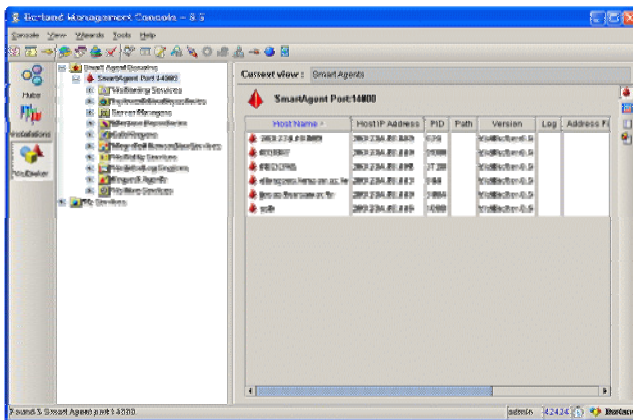


Fig. 6. System simulation environment presented in Borland management console

4.1 Results of Simulations

The simulation used 10 nodes for information processing and we calculate the processing time and accuracy of the information. In Figure 7 we present the result of processing time that has a filter variable set from 0.2 until 0.8. It is observed that when

increasing the filtering value, the processing time will also increase. In Figure 8 we present the result of the accuracy of the information by changing the filtering value in the same manner in Figure 7. To provide more accurate information and to process the information faster, we found the best value of filtering that has at least 80 percent accuracy of information and lesser processing time. Based on the graphical result, the filtering value of 0.5 has satisfied these conditions and will be used for comparison of the other algorithm.

The result of processing time by changing the filter value

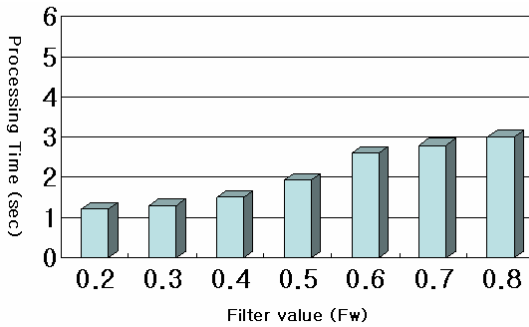


Fig. 7. Results of processing time

The result of accuracy of information by changing the filter value

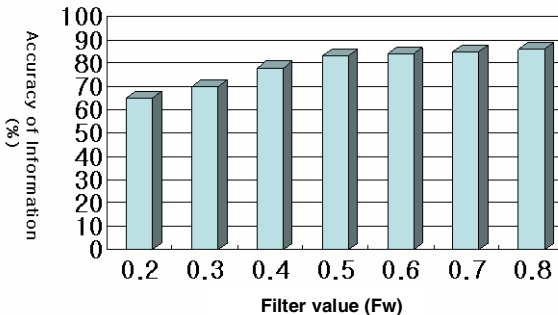


Fig. 8. Results of accuracy of information

We compared our proposed algorithm with SOM and F-NFC. We assumed that the number of nodes is the number of the multi-agents. The simulation used 100 nodes to process the information, setting up the learning rate to 0.5 and perform 100 times of looping value that was applied in all methods. We set the filtering value of the

FNC to 0.5. The result of the simulation is illustrated in Figure 9. The graphical results show that the FNC with filtering has the fastest processing time comparing with the three other methods. It was also observed that increasing the number of nodes will dramatically increase the processing time in SOM and F-NFC because it does not support filtering of data while NFC has lesser processing time because of filtering.

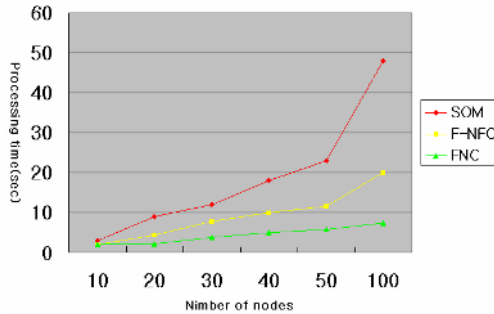


Fig. 9. Comparison result of the processing time by the methods

5 Conclusions

The multi-agent system provides transparency and accuracy service to the user through the interaction with multi-agent in the distributed environment. In addition to multi-agent interaction supports satisfaction of the users by providing more reliable information and faster information processing service.

In this paper, we used CORBA for the multi-agent interaction to solve the heterogeneity of the distributed environment. Also, we proposed a method of NFC with filtering that solved the problems of the existing research which has longer processing time using neural networks. The result implies that using our proposed algorithm can increase the reliability of the information processed and reduced the processing time of the search. The resource agent gathers resources in the distributed environment and performs the data mining. These data are pushed to the broker agent and it creates an agency which interacts with the multi-agent and process the data using the NFC with filtering method. All information is stored in the management table and the user can access the information needed from the management table.

The result of the proposed system simulation is evaluated over 80 percent accuracy of information, which is similar to other existing research but has reduced the processing time and is more effective in case the number of nodes increase.

This research uses an interaction of multi-agent and it applies to the information search system. The future works will be focused on the push service and management table.

References

1. Martín-Bautista, M., Vila, A., Sánchez, D., and Larsen, H.: Intelligent Filtering with Genetic Algorithms and Fuzzy Logic. *Technologies for Constructing Intelligent Systems*, (2002) pp. 351-362
2. Tu, H., and Hsiang, J.: An Architecture and Category Knowledge for Intelligent Information Retrieval Agents. *Proceeding of the Thirty-First Annual Hawaii International Conference on System Sciences*, Vol. 4. (1998) pp. 0405
3. Mukhopadhyay, S., Peng, S., Raje, R., Mostafa J., and Palakal, M.: Distributed Multi-agent Information Filtering - A Comparative Study, *Journal of the American Society for Information Science and Technology*, Vol. 56, Issue 8, pp. 834 - 842
4. Bhide, M.: Adaptive push-pull : Disseminating Dynamic Web Data, *IEEE Transaction on Computer*, Vol.51, No.6, (2002) pp.652-668
5. Ko J., Gerardo, B. D., Lee, J., and Hwang, J.: Information Search System Using Neural Network and Fuzzy Clustering Based on Mobile Agent, In the *Proceeding of ICCSA* (2005) pp. 205-214
6. Lee, J., Park, M., Gerardo, B., and Byun, S.: Reliable Information Search Mechanism through the Cooperation of Multiagent Systems in the Distributed Environment. *2nd ACIS International Conference on Software Engineering Research Management & Applications*, (2004) pp. 184-188
7. Nürnberger, A., Klose, A., and Kruse, R.: Self-Organising Maps for Interactive Search in Document Databases, *Intelligent Exploration of the Web* (2002)
8. Park, S., and Wu, C. Intelligent Search Agent for Software Components, *Sixth Asia-Pacific Software Engineering Conference* (1999) pp. 154
9. Carmine, C., d'Acierno, A., and Picariello, A.: An Intelligent Search Agent System for Semantic Information Retrieval on the Internet, *Proceedings of the 5th ACM International Workshop on Web Information and Data Management* (2003) pp. 111 – 117
10. Yong S. Choi and Suk I. Yoo.: Multi-agent learning approach to WWW Information Retrieval Using Neural Network, *Proceedings of the 4th International Conference on Intelligent User Interfaces*, (1998) pp. 23 – 30

Network Anomaly Behavior Detection Using an Adaptive Multiplex Detector

Misun Kim¹, Minsoo Kim², and JaeHyun Seo²

¹ Dept. of Computer Engineering, Mokpo Nat'l Univ.,
Mokpo, 534-729, Korea
misun@mokpo.ac.kr

² Dept. of Information Security, Mokpo Nat'l Univ.,
Mokpo, 534-729, Korea
{phoenix, jhseo}@mokpo.ac.kr

Abstract. Due to the diversified threat elements of resources and information in computer network system, the research on a biological immune system is becoming one way for network security. Inspired by adaptive immune system principles of artificial immune system, we proposed an anomaly detection algorithm using a multiplex detector. In this algorithm, the multiplex detector is created by applying negative selection, positive selection and clonal selection to detect anomaly behaviors in network. Also the multiplex detector gives an effective method and dynamic detection. In this paper, the detectors are classified by K-detector, memory detector, B-detector, and T-detector for achieving multi level detection. We apply this algorithm in intrusion detection and, to be sure, it has a good performance.

1 Introduction

The biological immune system is a mechanism of protecting itself by distinguishing foreign invasion material and clone generation and so on. Artificial Immune System (AIS) is a study field to apply the biological immune system to computer sciences [1]. In the biological immune system, a multi-level defensive mechanism makes a fast and adaptable response against many different kinds of foreign pathogens [2].

We propose an algorithm that efficiently detects and responds against network intrusions by utilizing the information processing capability of immune system for intrusion detection system. To detect an anomaly behavior, this algorithm uses multiplex detector (B-detector/T-detector) that is created by humoral immune response of B-cell and T-cell. The multiplex detector includes the detector (K-detector) acquired from well-known attack patterns in addition to B/T-detector which is generated from humoral immune response. The detectors are shown as an anomaly detector and a misuse detector because of using well-known attacks.

In existent anomaly detection system, we have difficulties in reflecting all the network data which is continuously changing while the first detector set was not changed [3]. In order to resolve these difficulties, we update on self-space by appending the normal data which is occurred through continuous on-line monitoring. Also, the clonal selection we use makes the detector update and generates new memory detector for effective detecting anomaly behaviors. The proposed algorithm is experimented with network dataset and the performance is shown through ROC curve.

2 Artificial Immune System

The biological immune system is composed by innate immune system and adaptive immune system [4]. The innate immune system exists irrespective of an antigen and reacts immediately to the foreign molecules. So, it takes priority of the elimination. The adaptive immune system is originated from generation of the antibody which is able to recognize an antigen. Then it activates the lymphocyte which recognizes special antigen. Then the lymphocyte removes an antigen through binding with that antigen. When antigens were not perfectly removed by innate immune system, it goes to the secondary defense system. The adaptive immune system has propensity to remember antigen that entered before, and it can remove antigen more effectively than innate immune system about antigen that invades often, because it reacts even faster than the first in the case of same antigen's entering.

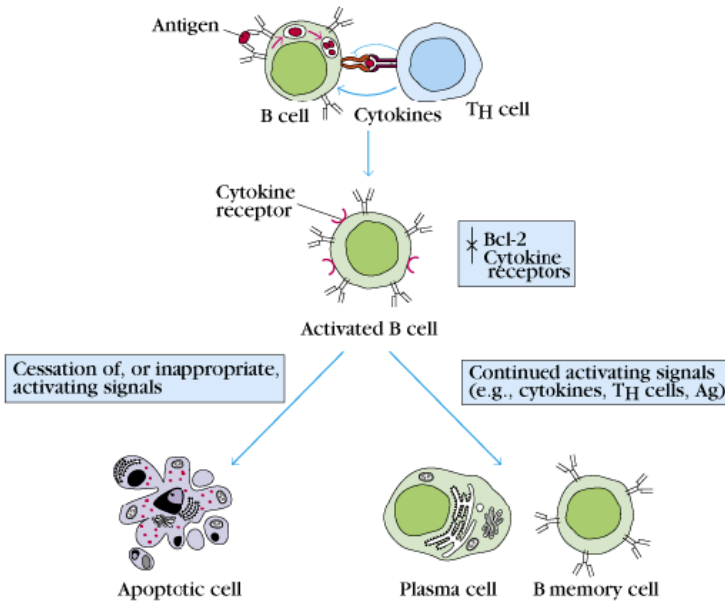


Fig. 1. The activation of B-cell and T-cell

Base element composing the adaptive immune system is B-cell and T-cell by lymphocyte of two forms. A B-cell produces and sends antibody that removes through binding on recognizing special antigen. A T-cell removes an antigen, controls growth of a B-cell, helps B-cell, or removes B-cell that do not recognize antigen normally. In adaptation immune systems, an antigen interacts with antibody through negative selection, positive selection, and clonal selection modules, and then it generates an immune answer. Fig.1 shows interaction process between antigen and lymphocyte cell [4].

The biological immune system uses a multilevel defense through the innate immune system and the adaptive immune system about invaders. The anomaly behavior

detection problem needs a multilevel detection technique with high detection rate and low false alarm rate like the immune system.

There are a lot of computer model fields based on an artificial immune system. Negative selection algorithm for anomaly behavior detection based on the discrimination between self and non-self is developed by Forrest [5]. Many researches used negative selection in many ways [6, 7, 8]. This algorithm generates detector randomly and removes the ones that detect self, so that the remaining detectors can detect non-self. If a behavior is detected by any detector, it is recognized as non-self.

We propose multiplex detector, and update self-space by appending the normal data that occurred through continuous on-line monitoring. Also, we continuously update detector through clonal selection, and generate memory detector, so that it may be effective for anomaly behavior detection. For an effective detector, we apply clonal selection theory of an artificial immune system. Clonal selection theory is able to respond quickly and efficiently to specific antigens [1]. First, it selects and activates molecules that are able to recognize a special antigen, and increases the number of clon. That is, it reproduces the cell which has ability of effective detection.

Because antigen does continuous change, antigen detection efficiency is kept by evolution process of B-cell antibody through clonal selection. Due to this memory effect, recognized antigen can be searched more quickly on the next invasion. By modeling clonal selection process, we produce a dynamic and efficient memory detector. Creating Dynamic and efficient memory detector makes it matches them to network data and detects anomaly behavior faster than any other detector. So we can reduce the seek time, and this dynamic detection mechanism drives more efficient anomaly detection.

3 Anomaly Detection Algorithm Using Multiplex Detector

Anomaly detection system proposed in this paper is to face an unknown attack. We use modeling self-discrimination function and clonal selection function of adaptive immune system. For effective and intellectual detection, anomaly detection method is often combined with misuse detection method [9]. To raise detection rate, we also use misuse detection mechanism so that it can detect a well-known attack. Fig. 2 shows anomaly detection system, which is consisted of preprocessing module, pattern generation module, detector generation module, detection module, and response module.

3.1 Proposed Anomaly Behavior Detection Algorithm

Anomaly behavior detection algorithm with multiplex detector based on network is following as Fig.3. We adopt clonal selection theory of the AIS for efficient detecting. Also it updates self-pattern by appending the normal data monitored on line, so it can react dynamically to the changing data set.

In this algorithm, detection processing is performed with an order of K-detector (detector of known attack), memory detector, and B/T detector for an input monitoring data. The result returns to the multiplex detector generation step, where the changed data set while the detection is performing can be detected. A multiple

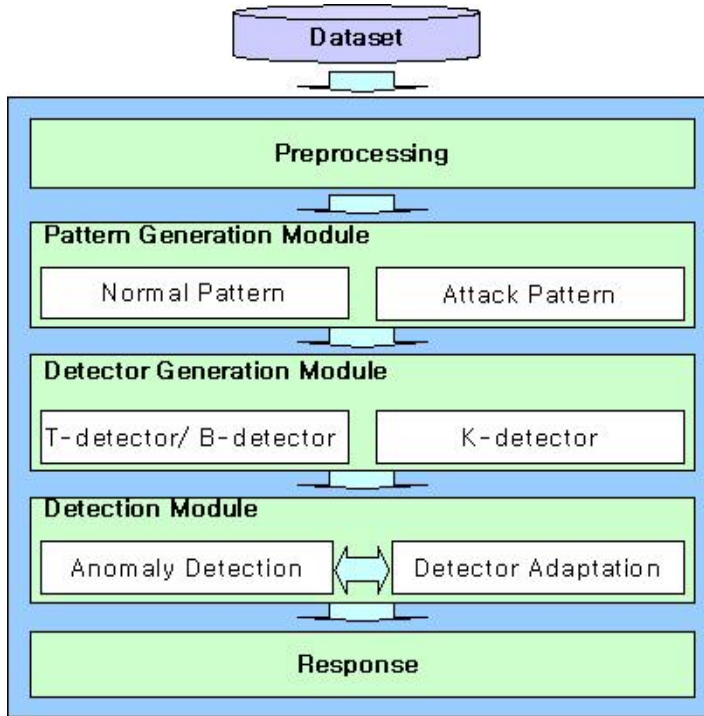


Fig. 2. The modules of anomaly detection using multiplex detector

detectors reduce the detection time, and also reduce the process time by automatic removing of detector which has low detecting rate for a certain period.

3.2 Function of Each Module

Preprocessing Module. Network based intrusion detection detects an intrusion by analyzing network traffic data. This module classifies the traffic data by network service using network packet header on TCP/IP and profiles normal behaviors by classified service to detect anomaly behaviors. Network normal behavior is constituted with the selected signatures such as network service number, the number of packets in a session, TCP connection, reconnection, simplex communication, the information of packet flags, and so on. In the preprocessing step, the data pattern is generated by extracting 12 signatures from network traffic data.

Pattern Generation Module. Pattern generation module generates both normal patterns and abnormal patterns. The normal patterns are generated from dataset including only normal behaviors while the abnormal patterns are from attack dataset. The both of two type patterns assume the form of binary string of data patterns

generated in preprocessing module. It is the reason of matching the packet data with continuous bits rules.

Detector Generation Module. This module creates the multiplex detector consisted of T-detector, B-detector, and K-detector, and they are from normal behavior patterns and attack patterns.

T-Detector Generation. In this step, the normal self-pattern is divided into slices with a fixed window size for matching r -contiguous bits. The mismatched data with the slices is selected to T-detector. The r -contiguous bit matching technique is a method that matches r -adjacent cells with patterns in part.

B-Detector Generation. B-cell receptor, in immune system, recognizes a certain part of antigen and binds it. This module draws first bit from bit stream of each signature in self pattern to generate a self-slice which has windows size 12. This module generates a B-detector from the self-slice by negative selection algorithm.

K-Detector Generation. K(nown)-detector is created from attack pattern, which is generated by pattern generation module. K-detector is used in misuse detection. This module creates the attack patterns from known attacks and draws the detector

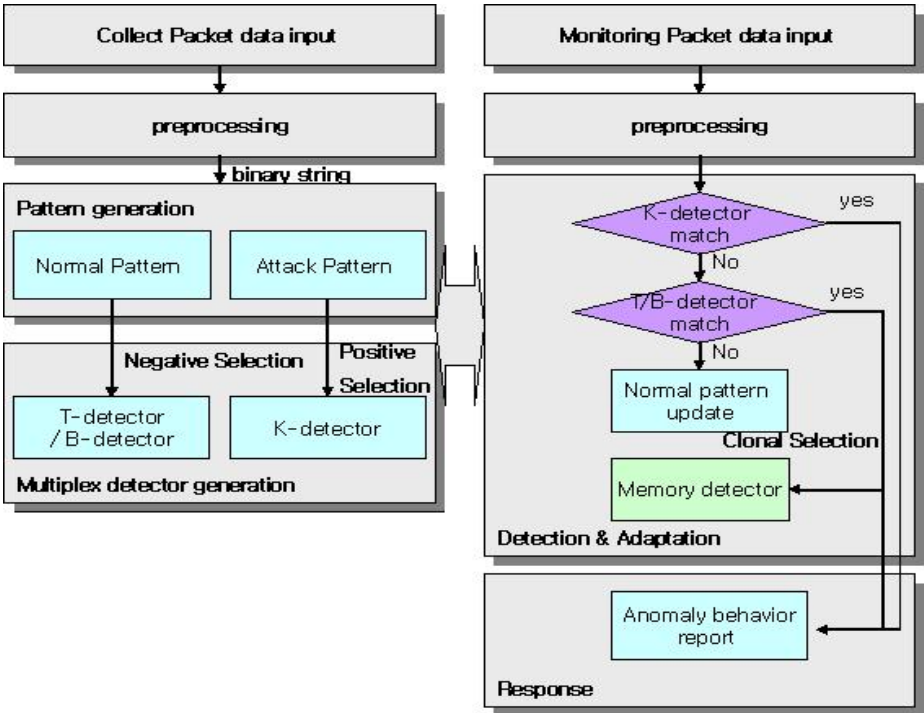


Fig. 3. Proposed Anomaly detection algorithm

from the patterns using positive selection techniques. This detector, namely K-detector, is used in initial detection of abnormal behaviors by matching it with monitored data.

Detection Module. In this module, anomaly behaviors are detected by matching the monitored data with multiplex detector created by detector generation module. If the monitored data is appended, at first, they are compared with K-detector to detect the anomaly behavior. If not, through the combination of T-detector and B-detector, they are distinguished whether they are the anomaly behavior or not. In this step, if a detector reacts on anomaly behavior, it becomes the memory cell. If a detector doesn't react for a certain period, it is deleted.

In this paper, the detector is classified into K-detector, memory detector, B-detector, and T-detector. The multi-level detection is required for a detection of multiplex detectors. K-detector is created through positive selection of attack pattern, and it is achieved the most preferentially in the detecting process. Because K-detector has information about known-attack, once detected by K-detector, it is considered as anomaly behavior. When it is not detected by K-detector, next step is detection by memory detector, and then by B/T-detector. When it is not detected by B/T-detector, it is not considered as anomaly behavior, so it updates self-set through self-pattern. B/T-detector increases the number of detection times when a reaction is occurred. If it reaches to a detecting response threshold predefined in system, it is changed to memory detector. If B/T-detector does not achieve certain times of detection for a certain period, then the B/T-detector is removed.

Response Module. Response module is a module that responds quickly to anomaly behavior which is detected in detection module. When anomaly behavior is detected, this module produces the alert signal.

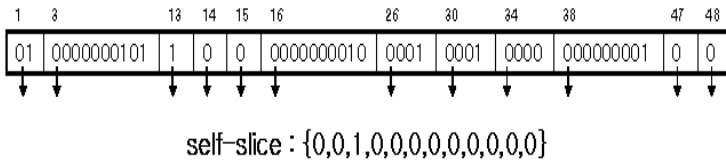


Fig. 4. Self-slice for generating B-detector

3.3 Algorithm Overview

In this algorithm, anomaly behavior detection processing for input monitoring data is performed in the following order: K-detector (detector of known attack), memory detector, B/T detector. The result is returned to the multiplex detector generation step, so the detection of changed data set can be possible.

This paper, due to the limit of page, does not include the describing entire algorithm in detail, but focuses instead on the processes of the detection module. The

process of anomaly detection and generating memory detectors performed by the detection module is as follows:

```

Input : K(K-detector-set), M(Memory-detector-set), B(B-
detector-set), T(T-detector-set), I(Input-data-set), dt
(Detection Threshold),lt(Life time threshold)
output : detect, N(Normal-data), TP, TN, FP, FN
begin
  dc = 0; // dc (B-detector's detection count)
  M=null;
  TP=0, TN=0; FP=0; FN=0;
  while(I){
    if(I==K)
      return detect=true;
    else if (M && I==M && I==T)
      return detect=true;
    else if (I==T && I==B)
      dc++;
      If (dt<=dc) M=M+B;//Memory-detector generate
      return detect=true;
    else
      L++; //L(B-detector's Life time)
      If (lt<=L) B--;//B-detector remove
      Return detect=false, N;
  }
  if (I is anomaly behavior)
    if (detect==true) TP++;
    else FN++;
  else if (I is normal behavior)
    if (detect==true) FP++;
    else TN++;
end

```

4 Simulation Results and Analysis

In this paper, we used a part of NT data in DARPA 2000[10] for simulation. Whole 2000 training data is sampled, and the attack pattern has 20 attack packets about ftp service. The first detector was created, where the string length of self-pattern is 48bit, the window size of B-detector is 12, window size of T-detector is 10, and the number of K-detector is 20. For evaluating the performance of multiplex-detector, we experimented on both case of single detector and multiplex detector. As the results, the detection rate is shown through ROC curve in Fig. 5.

As you see in Fig. 5, the multiplex detector has more high detection rate than the single detector. In general, the performance of anomaly detection system applying immune system is affected by the threshold of each detector. Therefore, it needs to find a reasonable threshold through experiments of the various thresholds about data set. The suitable threshold should have influence on performance of this algorithm.

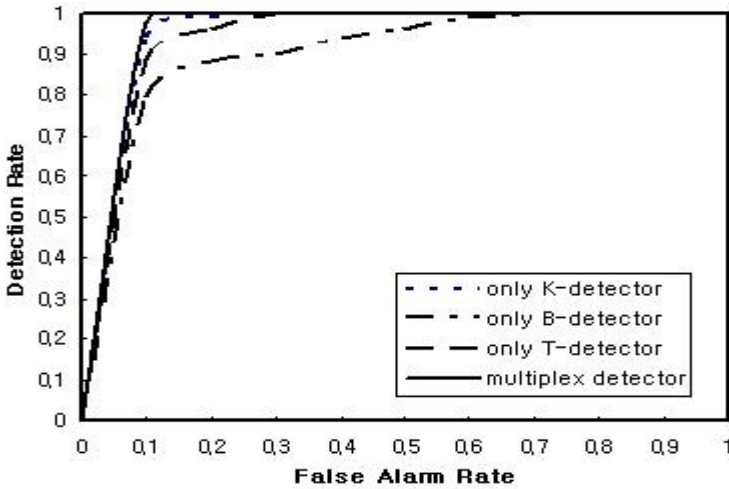


Fig. 5. Comparison of ROC curves with single detector and multiplex detector

5 Conclusions and Future Research

Anomaly detection algorithm proposed in this paper detected not only misuse detection but also anomaly detection, and multiplex detector brought better effect than single detection mechanism did.

Because existing anomaly detection system detects an anomaly behavior with a detector set which is made at the first time, there was difficulty to reflect network traffic data that is continuously changing according to an environment. So we made the data monitored continuously on line, append them to normal data set to keep changing self-pattern. This clonal selection method of continuous updating detectors and generating memory detectors makes the anomaly behavior detection more effective.

Anomaly detection system with immune system is under the effects of threshold setting for each detector. Therefore, experiment of various packet data group and threshold setting should be achieved. Then it can reduce detection time by optimizing algorithm, and provide efficient anomaly detection.

Acknowledgment

This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

References

1. de Castro, L. N. and Von Zuben, F. J, "Artificial Immune Systems: Part I – Basic Theory and Applications," Technical Report – RT DCA 01/99, 1999
2. Dasgupta, D., Yu, S. and Majumdar, N., "MILA – Multilevel Immune Learning Algorithm," GECCO 2003, LNCS 2723, pp.183-194, 2003

3. Chowdhury, D., "Immune Network: An Example of Complex Adaptive Systems," *Artificial Immune Systems and Their Applications*, 1st edition, Part II, Springer, pp.89-114, Dec. 1998
4. Goldsby, R., Kindt, T., and Osborne, B., *Kuby Immunology*, 4th Edition, W.H. Freeman & Company, Jan. 2000
5. Forrest, S., Perelson, A., Lawrence Allen, and Rajesh Cherukuri, "Self-Nonself Discrimination in a Computer," In *IEEE Symposium on Research in Security and Privacy*, pp.202-212, May 1994
6. Dasgupta, D. and Forrest, S., "An Anomaly Detection Algorithm Inspired by the Immune System," *Artificial Immune Systems and Their Applications*, 1st edition, Part III, Springer, pp.262-275, Dec. 1998
7. Kim, J. and Bentley, P., "Evaluating Negative Selection in an Artificial Immune System for Network Intrusion Detection," *Genetic and Evolutionary Computation Conference 2001 (GECCO-2001)*, San Francisco, pp.1330 - 1337, July 2001
8. Gonzalez, F. and Dasgupta, D., "Anomaly detection using real-valued negative selection," In special issue of the *Journal of Genetic Programming and Evolvable Machines*, pp 383-403, Vol. 4, Issue 4, Dec. 2003
9. Depren, O., Topallar, M., Anarim, E., and Ciliz, M.K., "An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks," *Expert Systems with Applications*, Vol. 29, Issue 4, pp.713-722, Nov. 2005
10. *DARPA Intrusion Detection Evaluation*, MIT Lincoln Laboratory, <http://www.ll.mit.edu/IST/ideval>

Applying Product Line to the Embedded Systems

Haeng-Kon Kim

Department of Computer Information & Communication Engineering,
Catholic University of Daegu, Korea
hangkon@cu.ac.kr

Abstract. For software intensive systems, a reuse-driven product line approach will potentially reduce time-to-market, and improve product quality while reducing uncertainty on cost and schedule estimates. Product lines raise reuse to the level of design frameworks, not simply code or component reuse. They capture commonality and adaptability, through domain and variability analyzes, to be able to create new products easily by instantiating prefabricated components, adapting their design parameters, and leveraging from established testing suites. In this paper, we examine software technology and infrastructure (process) supporting product lines more directly to embedded systems. We also present evaluation criteria for the development of a product line and give an overview of the current state of practices in the embedded software area. A product line architecture that brings about a balance between sub-domains and their most important properties is an investment that must be looked after. However, the sub-domains need flexibility to use, change and manage their own technologies, and evolve separately, but in a controlled way.

1 Introduction

Today the trend in computer-based products, such as cars and mobile phones, is shorter and shorter lifecycles. As a consequence, time spent on development of new products or new versions of a product must be reduced. One solution to this emerging problem is to *reuse* code and architectural solutions within a product family. Besides shortening development time, properly handled reuse will also improve the reliability since code is executed for longer time and in different contexts [1]. However, reuse is not trivial when applying it to real-time systems since both functional behavior as well as the temporal behavior must be considered.

We propose a design process suitable for developing product lines for real-time systems. The process starts in a requirement capturing phase where the requirements from all products in the line are collected. Communalities in functional- and temporal requirements among the products will be considered when the actual PLA is designed. The PLA is then analyzed. The objective of analyzing the PLA is to gain confidence in that the PLA is flexible enough to be a base on which all products can be realized without violating any temporal constraints. The contributions of this work with respect to embedded systems are an outline of a development process, focusing on the special considerations that must be taken into account when designing a PLA for embedded real-time systems. We also present an industrial case study of the use of a PLA for construction equipment.

2 Related Works

2.1 Principles of Software Product Lines

The software market - just like other markets - has a great demand for variety in products. The entirety of product-variants of one software is also referred to as software system family or software product line. Manufacturing was the first discipline that provided an answer how to efficiently build varying products. Instead of building single system family members, interchangeable parts were assembled to products. Actually the same principle can be transferred to software products, which is depicted in Figure 1. Reusable parts are identified & selected from a pool of reusable assets - the asset base - and integrated into single products.

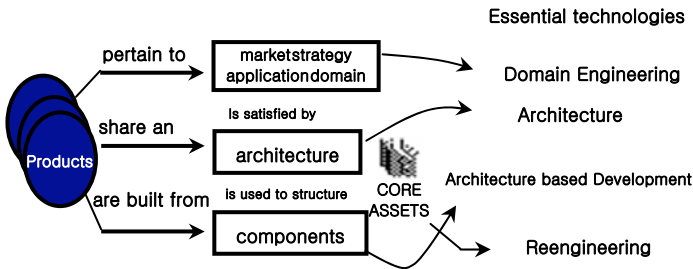


Fig. 1. Components in Product Line

2.2 Commonalities and Variabilities

Reuse implies that the asset base contains artifacts which can be reused in more than only one single product. This means that the asset base contains common parts which do not change between product line members and variable parts that feature different functionalities from member to member. Variability can be realized on run-time or development time. Product line engineering is concerned with development time variabilities as in figure 2.

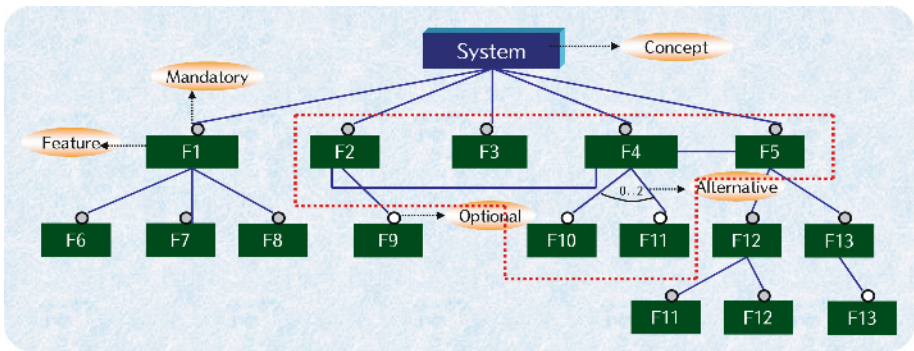


Fig. 2. Commonalities and Variabilities

Our focus in the subsequent part is on design techniques that can be applied for modelling product line variabilities (commonality does not require special design techniques). All presented techniques can be applied to components and work on an architectural level which means that they are independent from the later to be used implementation technology or programming language. However, there are differences in how good a given implementation technology or language may support these design techniques. Domain engineering consists of requirements analysis in embedded software domain, architecture analysis in embedded software and developments of assets for PL. Application engineering consists of requirements analysis in embedded software, system design and implementation.

Analysis of properties that are of vital importance for a PLA, e.g. flexibility. Moreover, the derivation of products from a PLA is dealt with. The development process proposed in this paper is shown in Figure 1 where the process is divided into *domain engineering*, *PLA development*, and *application engineering*.

3 Applying Product Line to Embedded Systems

3.1 Embedded Software Components

Components are pre-implemented software modules and treated as building blocks in integration. The integrated embedded software can be viewed as a collection of communicating reusable components. Figure 3 shows the embedded software constructed by integrating components. The component structure defines the required information for components to cooperate with others in a system. Our software component is modeled as a set of external interfaces with registration and mapping mechanisms, communication ports, control logic driver and service protocols, as shown in Figure 4.

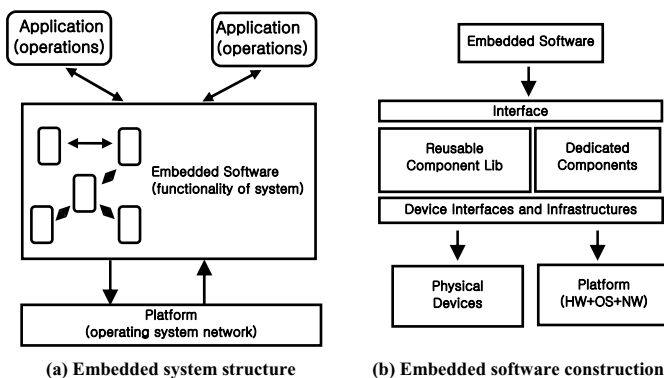


Fig. 3. Integration of embedded software components

3.2 Embedded Systems Development Process Using PL

Application engineering consists of requirements analysis in embedded software, system design and implementation. The achievement of Design-for-commonality and

control of-variability requires the establishment of a reuse infrastructure. There is a need for an overall framework that integrates the corresponding set of modeling, planning, and asset construction activities necessary for systematic reuse, and that, at the same time, allows the assimilation of technology effectively. There are three basic phases in the overall systematic reuse process: Domain Modeling, Product Line Design and Implementation, and product Development. In Figure 4 below, we illustrate the interrelations between these three phases. The primary information elements manipulated are in the Domain Knowledge $\{W\}$, which encompasses known facts about the domain environment or outside World. The domain modeling activity is constrained by the chosen scope $\{d\}$ thus sub-setting the domain knowledge to the specific set of products $\{Wd\}$ based on strategic goals and organizational mission. The Requirements $\{R\}$ define needs from an end-user point of view. The specification $\{S\}$ is produced as a precise description of $\{Wd\}$ from where an optimal Product $\{P\}$ can be built. The Target Platform, or Machine, $\{M\}$ provides the specific computing environment(s) on which the delivered products and assets will execute. One of the most critical questions is to be able to define correct, unambiguous and useful mappings between all these sets of conceptual elements and generated artifacts. The domain modeling process primarily focuses on domain analysis and system-level conceptual design to produce a generic problem description. The various representations of Wd and R contain models from different viewpoints. Each of these models highlights different aspects of the world, needs, and requirements, and are collectively referred to as *domain models*. Domain models describe typical systems features, including functional and non-functional, as well as mandatory and optional features.

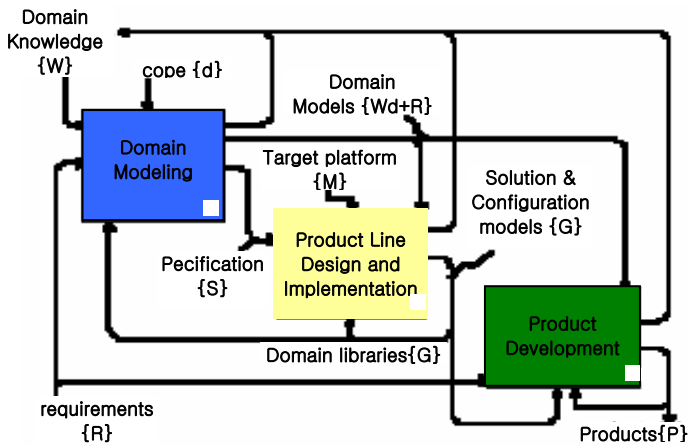


Fig. 4. Overall Embedded systems reuse process

3.3 Design and Implementation of Embedded Systems

Though many software product and process technologies are already available, the embedded software domain puts specific demands to the application of these technologies. Dedicated research results and products are present for software architecture

development and assessment, requirements engineering and validation, software process improvement, and tools to support all these technologies. However the major disadvantages of these technologies are that they do not take into account the specific needs for embedded systems and that they are applied “stand alone”, which in many cases is not very effective and leads to disappointing results. The embedded systems industry puts specific demands to the usage of such methodologies, such as the large dependency on real-time features, limited memory storage, large impact of hardware platform technology and the related cost drivers of the hardware, etc. The existing software engineering methodologies do not distinguish the specific impacts or necessary customization for the embedded domain, nor is it indicated how they should be used specifically for each specific area within this domain, i.e. automotive, telecom, consumer electronics, safety critical, etc. The embedded software domain puts dedicated pressure on these methodologies. Reasons for this are the high complexity of these products and the dependency in this domain on innovative highly technical solutions. Furthermore, the embedded domain is much more driven by reliability, cost and time-to-market demands. This makes the embedded domain a specific area for which available generic methodologies need to be adapted.

We see a variety of implementation mechanisms as required to enable real-time and embedded software to safely interact with software written for more mainstream applications. We assume that these two kinds of software will share objects and interact by performing atomic operations that update the shared objects.

The integration of existing software engineering technologies into one workbenches is required, in order to achieve an effective overall approach, which is

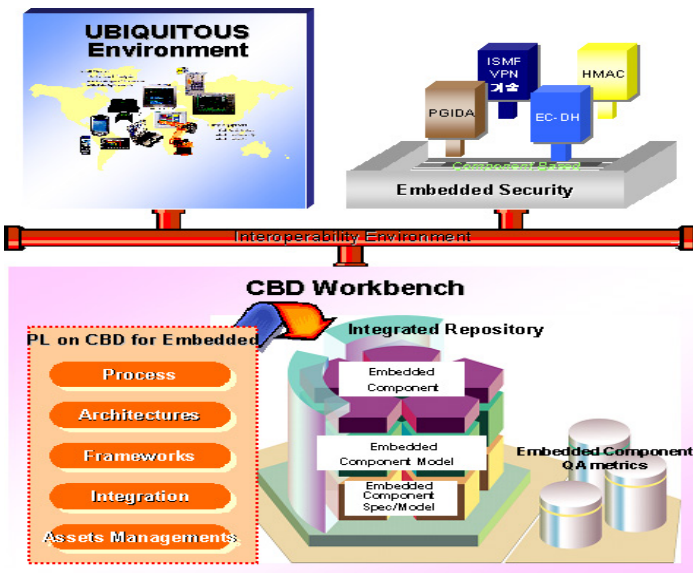


Fig. 5. Embedded Systems Developments Workbenches

customizable to the needs of each specific software development of embedded systems as in figure 5. Application of these methodologies requires mutual information exchange and co-operation along specific interfaces. The workbenches will arrange these interfaces and will make it possible to exchange methodologies, depending on the specific needs of the product and the organization, as if they are components. The framework contains: processes, methods, techniques, tools and templates supports the decision making model. The fact is that there are too many methods or approaches available and, therefore, there is a need to reduce uncertainty in the decision making in selecting the set of methods, techniques and tools for a certain use situation. That means that support for decision making in choosing the method for a certain situation is needed. Quite typical situation where the framework is needed is a project aiming to choose the most effective and efficient way for the development of embedded systems. The framework integrates knowledge about processes, methods, techniques, tools and templates into a decision making model .

3.4 Product Development

The PL design and implementation phase generates generic *solutions* and *configuration* models with specific information to support adaptation. **Solution models** represent both software and hardware architectures (i.e., components and their interfaces) suitable for solving typical problems in the domain. The PL design activity fine-tunes the system-level architectural design with detailed common and flexible product and component structures. These detailed designs correspond to *product line architectures* and *component designs*, respectively, documented as *solution or design models*.

Product line architectures depicts the structure for the design of related products and provide models for integrating optional/alternative components. Component Designs specify the structure for the explicit variability of components across products and product lines; they serve as models for specifying and encapsulating commonality.

Optional parts are very important. Possible combinations of optional features must be supported by the product line architectural design and the flexibility incorporated in the component design, otherwise it may be impossible to reuse an asset “as-is” because commonality and specificity are mixed. This would make it necessary to modify the asset when reused in the new context, and this should obviously be voided whenever feasible. Configuration models help here.

A *configuration model* maps between the problem models and solution models in terms of product construction rules, which translate capabilities into implementation components. These rules describe legal *feature* combinations, default settings, etc. For example, certain combinations of features may be not allowed; also, if a product does specify certain features, some reasonable defaults may be assumed and other defaults

The Product development process is multiplexed to correspond to the various product lines. Generative application development can be applied, whereby the application engineer states requirements in abstract terms, from where application generators produce the desired system or component. The actual development process follows any established development framework.

4 Systems Integration

Software integration includes component selection and binding, and control plan construction (both control logic and operation sequence). A runtime system can be generated by mapping the integrated software onto a platform as in figure 6.

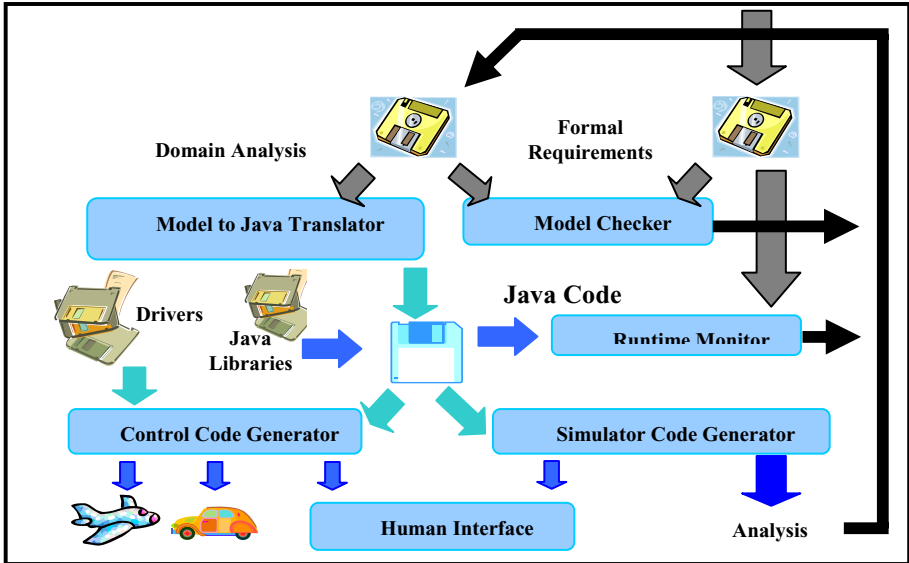


Fig. 6. Embedded Systems Integration

4.1 Composition Model

The composition model defines how software can be integrated with given components. Since each reusable component is implemented with a set of external interfaces that uniquely define its functionality, components can be selected based on the match of their interfaces and design specifications. The integration of reusable components can be viewed as linking the components in integrated can be viewed as linking the components with their external interfaces.

Reusable components in integrated software are organized hierarchically to support integration with different granularities, as illustrated in Figure 7. The behavior of an integrated component can then be modeled as integration of its member component behaviors. The control logic and operation sequences of each component can be determined individually and specified in a *Control Plan*. The device-independent behaviors depend only on the application level control logic, and can be reused for the same application with different devices. The device-dependent behaviors are dedicated to a device or a configuration, and can be reused for different applications with the same device.

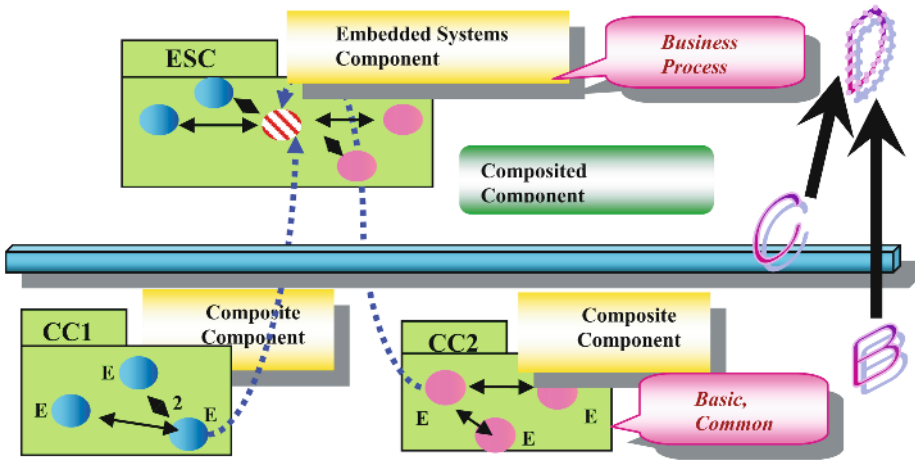


Fig. 7. Hierarchical composition model

With such a composition model, both components for low-level control such as algorithms and drivers and for high-level systems can be constructed and reused. However, additional overhead is introduced as the component level is increased, and may result in associated performance penalties due to excessive communications and code size.

5 Conclusion

Product line development aims at the reuse of software requirements, architecture, components and processes. Analysis of the commonality and variability of functional and structural properties of systems is the main difference between the development of a product line and the development of a single software system. In this paper, we have been formed in order to solve a challenging, complex, but highly relevant issue: integration between software engineering technologies for embedded systems. We also presented a component-based architecture for embedded software integration. This architecture defines components and a composition model as well as a behavior model. A reusable component in our architecture is modeled with a set of events as external interfaces, communication ports for connections, a control logic driver for separate behavior specification and reconfiguration, and service protocols for executing environment adaptation. Such a structure enables multi-granularity and vendor-neutral component integration, as well as behavior reconfiguration.

References

1. B. P. Douglass. Real-time UML: developing efficient objects for embedded systems. Addison-Wesley, 2000.
2. J. O. Ostroff. Formal methods for the specification and design of real-time safety critical systems. <http://www.cs.yorku.ca/jonathan/survey/combined-paper.html>.

3. Robertson, D., Ulrich, K. Planning For product platforms. *Sloan Management Review* Vol.39, No. 4, pp. 19-31 2001.
4. Schmid, K. Scoping software product line. An analysis of an emerging technology. In *Proceedings of Software Product Line. Experience and research directions*. Donohoe, P. (ed.), Kluwer Academic Publishers, Massachusetts, pp.513-532,2002.
5. Byeongdo Kang., Young-Jik Kwon., Lee, R.Y. A design and test technique for embedded software. *Software Engineering Research, Management and Applications*, 2005. Third ACIS International Conference on. Pp.160-165, 2005
6. Baleani, M., Ferrari, A., Mangeruca, L., Sangiovanni-Vincentelli, A.L., Freund, U., Schlenker, E., Wolff, H.-J., Correct-by-construction transformations across design environments for model-based embedded software development. *Design, Automation and Test in Europe*, 2005. *Proceedings*, Vol. 2, pp 1044-1049, 2005
7. Sikun Li., Zhihui Xiong., Tiejun Li., Distributed cooperative design method and environment for embedded system, *Computer Supported Cooperative Work in Design*, 2005. *Proceedings of the Ninth International Conference on*, vol. 2, pp 956-960, 2005
8. Qiao, Y., Berzins, V., LuqiQiao, Y., Berzins, V., Luqi., FCD: a framework for compositional development in open embedded systems, *Information Technology: Coding and Computing*, 2005. *ITCC 2005. International Conference on*, vol. 2, pp 479-484, 2005
9. Ramamritham, K., Arya, K., Fohler, G., System software for embedded applications, *VLSI Design*, 2004. *Proceedings. 17th International Conference on*, pp 12-14, 2004

Enhanced Fuzzy Single Layer Learning Algorithm Using Automatic Tuning of Threshold

Kwang-Baek Kim¹, Byung-Kwan Lee², and Soon-Ho Kim³

¹ Dept. of Computer Engineering, Silla University, Korea
gbkim@silla.ac.kr

² Dept. of Computer Engineering, Kwandong University, Korea
bklee@kwandong.ac.kr

³ Dept. of Automotive Mechanical Engineering, Silla University, Korea
skim@silla.ac.kr

Abstract. In this paper, we proposed an enhanced fuzzy single layer learning algorithm using the dynamic adjustment of threshold. For performance evaluation, the proposed method was applied to the XOR problem, which is used as a benchmark in the field of pattern recognition, and the recognition of digital image in a practical image processing application. As a result of experiment, though the method does not always guarantee the convergence, it shows the improved learning time and the high convergence rate.

1 Introduction

The conventional single layer perceptron is inappropriate to use when a decision boundary used to classify input patterns does not composed of hyper plane. Moreover, the single layer perceptron, due to its use of unit function, is highly sensitive to the change of weights, difficult to implement and can not learn from past data[1]. Therefore, it can not find a solution for the exclusive OR problem being a benchmark.

There are a lot of endeavor to implement a fuzzy theory to artificial neural network[2]. Goh et al.[3] proposed the fuzzy single layer perceptron algorithm and the advanced fuzzy perceptron based on the generalized delta rule to solve the XOR problem and the classical problems[3]. The single layer perceptron algorithm guarantees some degree of stability and convergence in application using fuzzy data, but on the other hand, it causes a considerable amount of computation and some difficulties in application to the complicated image recognition. Also, the enhanced fuzzy perceptron has shortcomings such as the possibility of falling in local minima and slow learning time[4][5].

In this paper, we proposed an enhanced fuzzy single layer learning algorithm using the dynamic adjustment of threshold. We constructed, and trained, a novel single layer perceptron using the auto-tuning method of threshold. And through performance evaluation, we showed that the proposed method can guarantee to find solutions for problems such as exclusive OR and digit image recognition which the conventional fuzzy single layer perceptron cannot solve.

2 A Fuzzy Single Layer Perceptron

Wang[3] proposed a fuzzy single layer perceptron using the learning algorithm based on the generalized delta rule. This algorithm guarantees some degree of stability and convergence in an application using fuzzy data. However, it causes a considerable amount of computation and some difficulties in an application of complicated pattern recognition. Fig.1 shows the architecture of the fuzzy single layer perceptron.

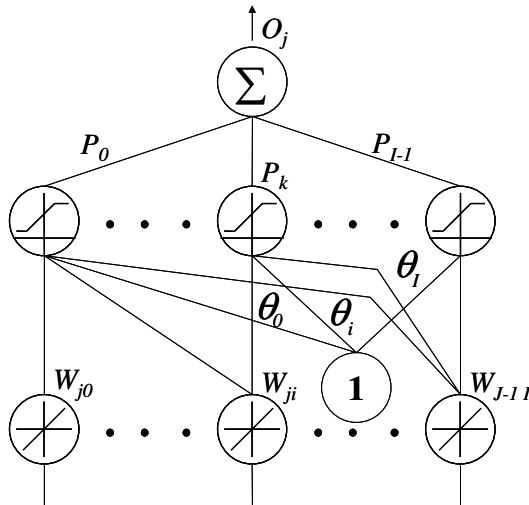


Fig. 1. A fuzzy single layer perceptron model

The fuzzy single layer perceptron can be simplified into four steps. For each input, it repeats all steps from step 1 to step 4 until error is minimized.

Step 1: Initialize weight and bias term.

Define W_{ji} to be the connection weight from input i to output j at time t , and θ_j to be the bias term in the output soma. Set $W_{ji}(0)$ to small random values for the initialization of all weights and the bias term.

Step 2: Rearrange A_i in the ascending order of membership degree m_i and add an item m_0 at the beginning of this sequence.

$$0.0 = m_0 \leq m_1 \leq \dots \leq m_i \leq m_l \leq 1.0$$

Compute differences between two neighboring items of the sequence.

$$P_k = m_i - m_{i-1}, \text{ where } k = 0, \dots, n$$

Step 3: Calculate the actual output of a soma (O_j).

$$O_j = \sum_{k=0}^{I-1} P_k \times f \left(\sum_{i=k}^{I-1} W_{ji} + \theta_j \right)$$

where $f\left(\sum_{i=k}^{I-1} W_{ji} + \theta_i\right) = \begin{cases} \mathbf{0} & \text{if } \sum_{j=k}^{J-1} W_{ij} + \theta_i < T, \text{ and } T \text{ is threshold.} \\ \mathbf{1} & \text{Otherwise} \end{cases}$

Step 4: Applying the delta rule, derive the gradually increasing changes for weight and bias term.

$$\Delta W_{ji}(t+1) = \eta_j \times E_j \times \sum_{k=0}^{I-1} P_k \times f\left(\sum_{i=k}^{I-1} W_{ji} + \theta_j\right) + \alpha_j \times \Delta W_{ji}(t)$$

$$W_{ji}(t+1) = W_{ji}(t) + \Delta W_{ji}(t+1)$$

$$\Delta \theta_j(t+1) = \eta_j \times E_j \times f\left(\sum_{i=k}^{I-1} W_{ji} + \theta_j\right) + \alpha_j \times \Delta \theta_j(t)$$

$$\theta_j(t+1) = \theta_j(t) + \Delta \theta_j(t+1)$$

where η_i is a learning rate and α_i is a momentum.

3 Automatic Tuning Method of Threshold

In the fuzzy set theory, many operators that aggregate the membership function were defined and used. The fuzzy intersection operator has the property that the output value is not greater than the minimum value among all input values, and Yager’s intersection operator[6] is described like Eq. (1).

$$\mu_{X_1 \cap X_2} = 1 - \min\left\{1, \left(\left(1 - \mu_{X_1}\right)^p + \left(1 - \mu_{X_2}\right)^p\right)^{\frac{1}{p}}\right\} \tag{1}$$

Yager’s intersection operator converges to the min-operator if $p \rightarrow \infty$. And the operator shows the special feature that the aggregated value is not greater than the smallest value among all inputs.

By controlling the operator and the parameter value needed by the transfer function of hierarchical neural network that uses such an operator, the pessimistic and optimistic propensity appearing in information merging can be decided automatically. In the problem of adjusting threshold of transfer function, the learning is performed with threshold automatically decreased from the maximum value “1” by fuzzy operator, i.e. Yager’s intersection operator satisfying each pattern simultaneously. Threshold is set by the compensative operator, which uses the calculated maximum and minimum value. In the case of this method, the network is prevented from a rapid decrement and becomes modified with satisfying automatically the pessimistic and optimistic degree.

In the conventional method, because each pattern is learned with a fixed threshold, it may result in an inappropriate classifying by decision plain. Therefore, we proposed an enhanced method that can classify the decision plain by controlling the threshold. That is, in the proposed method, the problem of fixed threshold was improved by the dynamic adjustment using the Yager’s generalized intersection operator. In the Yager’s generalized intersection operator, $p = 2$ was used on the basis of experiment. Fig.2 shows the proposed dynamic adjustment method of threshold.

$$\begin{array}{l}
 \text{Case of Comparison } (Error^{old} \& Error^{present}) \{ \\
 \quad " < " : \{ value^{max} = Thershold^{old} \\
 \quad \quad \quad value^{min} = Thershold \} \\
 \quad " = " : \text{Skip} \\
 \quad " > " : \{ value^{min} = Thershold^{old} \\
 \quad \quad \quad value^{max} = Thershold \} \\
 \} \\
 Error^{old} = Error^{present} \\
 Thershod^{old} = value^{max} \\
 Thershod = 1 - Min \left[1, \left((1 - value^{max})^p + (1 - value^{min})^p \right)^{\frac{1}{p}} \right]
 \end{array}$$

Fig. 2. The proposed dynamic adjustment method of threshold

4 Simulation and Result

We simulated the proposed method on IBM PC/586 with VC++ programming language. In order to evaluate the proposed method, we applied it to the exclusive OR, which is used as a benchmark in neural network, and the ID code pattern recognition problem, which is a kind of image recognition. In the proposed method, the error criterion was set to 0.05.

4.1 Exclusive OR

In this experiment, we set the initial values of learning rate and momentum to 0.5 and 0.75, respectively. Also we set the range of weight to [0,1]. In general, the range of weights were [-0.5,0.5] or [-1,1]. As shown in Table 1, the proposed method showed higher performance than fuzzy perceptron in convergence epochs and convergence rates of the three tasks. Fig.3 is the graph showing the change of threshold by the number of epoch. As shown in Fig.3, The threshold value was decreased automatically using fuzzy intersection operator.

Table 1. Convergence rate in initial weight range

Epoch No.	Conventional Fuzzy Single Layer Perceptron	Proposed Algorithm	Initial Weight Range
Exclusive OR	8	3	[0.0, 1.0]
	15	8	[0.0, 5.0]

The input units have an image pattern composed of 10×10 array of binary values. In the experiment, the conventional fuzzy perceptron was not converged, but the proposed method was converged on 63 steps on the image patterns. Table 2 is shown the summary of the results in training epochs between two algorithms.

Table 2. The comparison of epoch number

<i>Image Pattern</i>	<i>Epoch Number</i>
Conventional fuzzy single layer perceptron	0 (not converge)
Proposed algorithm	63 (converge)

5 Conclusions

We have proposed an enhanced fuzzy single layer perceptron using auto-tuning method of threshold, which has greater stability and functional varieties compared with the conventional fuzzy single layer perceptron. The proposed network is able to extend to the arbitrary layers and has high convergence in the case of two layers or more. In this paper, we considered only the case of single layer and the network showed a processing of high speed in the learning and a satisfactory result of recognition on huge image patterns. The proposed algorithm shows the possibility of application to image recognition as well as the benchmark test in neural network by single layer structure.

References

1. Judith, E., D.: Neural Network Architectures An Introduction. Van Nostrand Reinhold, New York (1990)
2. Gupta, M. M. and Qi, J.: On Fuzzy Neuron Models. Proceedings of IJCNN, Vol. 2. (1991) 431-435
3. Goh, T. H., Wang, P. Z. and Lui, H. C.: Learning Algorithm for Enhanced Fuzzy Perceptron. Proceedings of IJCNN, Vol. 2. (1992) 435-440
4. Kim, K. B. and Cha, E. Y.: A New Single Layer Perceptron using Fuzzy Neural Controller. In: Jaime, O., Ariel, S. (eds.): Simulators International XII, Vol. 27. No. 3. (1995) 341-343
5. Kim, K. B., Kim, S., Joo, Y. and Oh, A. S.: Enhanced Fuzzy Single Layer Perceptron. In: Jun, W., Xiaofeng, L., Zhang, Y. (eds.): Advances in Neural Networks – ISNN 2005. Lecture Notes in Computer Science, Vol. 3496. Springer-Verlag, Berlin Heidelberg New York (2005) 603-608
6. Kim, K. B., Seo, C. J. and Yang, H. K.: A Biological Fuzzy Multilayer Perceptron Algorithm. Journal of KIMICS, Vol. 1. No. 1. (2003) 99-103
7. Kim, T. K., Yun, H. G., Lho, Y. W. and Kim, K. B.: An Educational Matters Administration System on the Web by Using Image Recognition. Proceedings of Korea Intelligent Information Systems, (2002) 203-209

Optimization of Location Management in the Distributed Location-Based Services Using Collaborative Agents*

Romeo Mark A. Mateo, Jaewan Lee, and Hyunho Yang

School of Electronic and Information Engineering, Kunsan National University,
68 Miryong-dong, Kunsan, Chonbuk 573-701, South Korea
{rmmateo, jwlee, hhyang}@kunsan.ac.kr

Abstract. Determining mobility patterns and predicting the future mobile location are some issues in optimizing the location management of location based services. These patterns are also useful for finding trends, allocating resources and other information with the mobile user. In our paper, a location management on location based services using collaborative agents is presented. In this study, we propose an agent which learns from the pattern of the user mobility and predict the future location of mobile object to optimize the search method of the location management. We use the association rule mining and basing on these rules, the agent predicts the future mobile location. Also, these agents coordinate to each other by telling the knowledge of the possible location of the mobile object on the distributed location-based services. The result using the technique optimizes the search method of the location management.

1 Introduction

Nowadays, we can already acquire information services at any place and in any time by using mobile and ubiquitous devices. In cellular phones, we have services like banking, city guides, games and many more. Ubiquitous devices like smart cards provide services and awareness which have less resources but more powerful to provide information like location monitoring because of its availability. These services are provided by the location-based service [1], [2]. Also, information about the moving object is important and not all location services have the same database access schemes. Different location services simply mean different deployment of the databases. Using software agent [3] solves the problem in the distributed location services. The agent shares its data with the other parts of the system so that information like mobile location is visible to other location service.

In addition, the optimization of search and update method are topics of researches in location management. One of an optimization scheme is prediction of mobile location. A data mining approach about sequential mining of patterns of the mobile user movement in a certain region to predict the next inter-cell movement of the mobile object is presented [4].

* This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessments) (IITA-2005-(C1090-0501-0022)).

We propose a collaborative solution to the location management. The location management of the proposed system presents a hierarchical process of searching and updating and using agents to accomplish the task. The agent extracts mobility patterns of the user from the previous mobile request. The agent predicts the next location of the mobile object by using these patterns. The collaboration is done by consulting the agent for the knowledge of the mobile object's possible location. If the agent has no knowledge about the location then the collaboration is initiated. The design for the extraction of the user mobility pattern was on a modular form so that different data mining algorithms can be applied. In our research, we use the apriori algorithm to extract the mobility patterns of the user and based the next location of the moving object. These agents were deployed through CORBA.

The rest of the paper is organized as follows. In Section 2, we discuss some related works of location management. Section 3 explains the architecture of the proposed multi-agent location management. Section 4 presents the performance analysis of the optimized search method using location agent manager by analyzing the previous mobile request data to redirect the possible location of the mobile object and the simulation results in Section 5. Section 6 concludes the result and discusses the additional future works of the study.

2 Related Works

The main issues of the location management is the operation of search and update method of the mobile objects. Researches focus on improving these methods and considering the trade-offs obtained by the two operations. The next subsection explains the related studies in optimizing the location management.

2.1 Cellular Networks

Many research works tackle the methods of locating a mobile object in cellular networks [5] [6] [7]. Some introduce schemes to minimize the cost of paging and updating of location management. Research on location area schemes [8] discussed how the mobile station updates its location whenever it moves into a cell which belongs to new location area. Optimal sequential paging [5] presents a polynomial-time algorithm to solve the problem of minimizing the average paging costs. A predictive distance-based mobile tracking scheme [6] for prediction of mobile location and a dynamic location update scheme in velocity-based scheme [7] are presented.

2.2 Hierarchical Schemes

Hierarchical scheme of search and update method is another approach for location management discussed by Pituora, et. al. [9]. Using the hierarchical scheme leads to reduction of communication cost when the call is often done in local or regional area. The study uses a hierarchical distributed location database to track mobile object in which the nodes are networked in a tree-like structure. It extends two-tier schemes by maintaining a hierarchy of location databases. A location database at a leaf serves a single logical-cell (l-cell) and contains entries for all objects currently in it. A database at an internal node maintains information about objects residing in the set of l-cells in its sub tree. For each mobile object, the information becomes a pointer to an

entry at a lower database. In addition, a research of hierarchy of agents in the network nodes that acts as mobile location databases [10] can be a model for location management using agents.

2.3 Mobility Patterns and Prediction

Mobility prediction can optimize the location management by using location prediction of the mobile objects. An adaptive location prediction [11] uses a hierarchy of location area to estimates the location probabilities of each mobile user. A sequential mining approach for the location prediction is used to allocate resources in a PCS network [4]. There are three phases presented. First is mining the user mobility pattern, second is extracting the mobility rules and the last is the mobility prediction. The predicted movement can then be used to increase the efficiency of location management. Also, using this technique can effectively allocate resources to the most probable-to-move cells instead of blindly allocating excessive resources in the cell-neighborhood of a mobile-user. Figure 1 shows an example coverage region ($C_0 - C_8$) use for user mobility pattern.

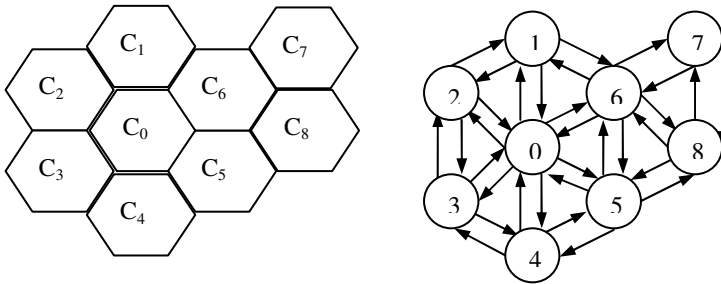


Fig. 1. Mining user mobility patterns from graph traversal

3 Location Management Using Collaborative Agents

The collaborative agents' goal was to optimize the search method of the location management by telling the agent's knowledge from the data of the mobile object. The design of the proposed system is compliance of the Common Object Request Architecture (CORBA). Figure 2 shows the implementation of location management using collaborative agents. The multi-agent component communicates through the Internet Inter ORB Protocol (IIOP). The CORBA [12] [13] is used to act as a middleware on the distributed environment. In our architecture, different location-based services use the ORB core to interoperate their services to each other. The services are managed by the location agent manager (LAM). Figure 2 shows each LAM has communication to other LAM by location agent.

Each LBS has a LAM which accesses the database and communicates with other agents like the location agents (LA) and collector agents. The collector agent updates and deletes the mobile object collected from the embedded systems. Location agents are the medium of communication of each LAM in LBS.

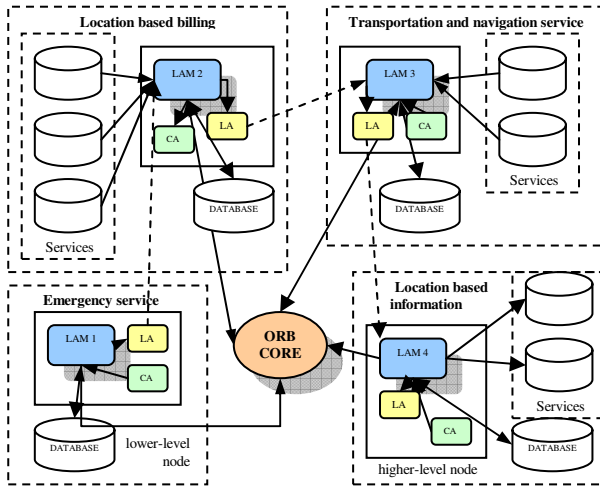


Fig. 2. Architecture of the distributed LBS implemented in CORBA

Figure 2 also shows the process of the proposed distributed location-based services in hierarchical search and update. First, mobile objects are stored in the database of LBS that are collected by the collector agent. The location agent manager 1 (LAM 1), that resides in the emergency service, copies the mobile object information and passes the value to be stored to LAM 2. This becomes a pointer of the original location of mobile objects. The update method continues until it reaches the root node. Each mobile ID from the lower-level passes the value to higher-level location agent. Next, the search has the same method of hierarchy. In addition to it, we developed the optimization algorithm. The proposed algorithm locates the possible location of the mobile object through collaborations. If the agent did not find the requested mobile object in the searched node then it communicates with the high-level agent if it knows the possible location of the mobile object.

3.1 Multi-agent Components

One of the components of the multi-agent in the distributed location based service architecture is the location agent and its task is to communicate with the parent agent on the hierarchical structure of the location agents. Figure 2 presents the hierarchical update method of the location agent. The mobile ID information is gathered by the lower-level agents and updates the value to the higher-level agent. Figure 3, the location agent is illustrated as a medium of communication between each node.

The main functions of location services are to collect and process the mobile objects to serve the mobile user subscribers. Active badge locating system [14] is one example of location service which provides information of staff members wearing the active badge within the establishment. The proposed system uses collector agents to collect data or mobile objects. These agents communicate with the location agent manager to update mobile objects information. In Figure 3, the collector agent is

presented between the input device and location agent manager. All input data like mobile object is processed to the collector agent and send it to the LAM.

The main component of the multi-agent is the location agent manager or LAM. The LAM functions are to access the database of LBS, manage the services of each node and communicate to other agents through location agent within the LBS. Figure 3 shows the LAM as the main component of the proposed system. Also the optimal search algorithm is implemented in this component. The LAM analyzes the previous mobile request to locate its location.

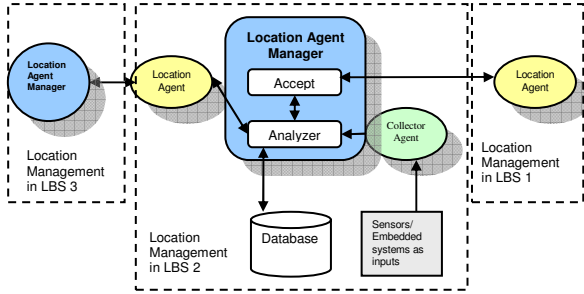


Fig. 3. Multi-agent components

3.2 Optimal Search Module

The optimal search method was used in the hierarchical location management. Research study using data mining [4] for user mobility pattern to predict the movement of the mobile objects that optimizes the search method. It presents a sequential mining of patterns in the cell to cell movement. In our proposed system, agent mines from the previous requests of a mobile object by using the Apriori algorithm so that it can provide an optimal search. Its function includes 3 phases. In the first phase, the previous data requests of mobile object are retrieved, collected and preprocessed by deleting the attributes that has a missing value. This preprocessing prepares it for data mining stage. Other preprocessing is also available like filling up the missing values of the attributes but this method will consume more time to process and was not considered to our work. Equation 1 represents the function of collecting and preprocessing of the mobile requests data $rec_i(t_1(node), t_2(node), \dots, t_x(node))$, x is indicated as the time sequence and i is the mobile identification.

$$C = \sum rec_i(t_1(node), t_2(node) \dots t_x(node)) \tag{1}$$

where $t_x(node) \diamond null$ value

The second phase is using the optimal search module which we can insert the method of pattern mining. We used the Apriori algorithm to generate the user mobility pattern on C . Apriori employs an iterative approach known as level-wise approach. The use of this algorithm is discussed by Gerardo et. al. in [14] where a proposed agent was used to extract association rules. Also, the research introduces a method on

minimizing the operation time by clustering prior to pattern discovery. Our algorithm which mines the mobility patterns also followed the two-step process for the Apriori.

1. Join step – find L_k , a set of candidate k -itemsets by joining L_{k-1} with itself.
2. Prune step – C_k is generated as superset of L_k , that is, its members may or may not be frequent, but all of the frequent k -itemsets are included in C_k .

The Apriori property implies that any $(k-1)$ -item that is not frequent cannot be a subset of a frequent k -itemset; hence, the candidate can be removed. After producing the patterns, mobility rules will be generated. In the last phase, these rules are filtered so that the search for the matching pattern from the recent existing patterns of the user becomes reliable and efficient.

3.3 Collaboration of Agents

In our research, collaboration is the process of consulting the agent about its knowledge of the previous mobile location. The collaboration is initiated if the source agent has no knowledge about the probable location of mobile object that is needed for

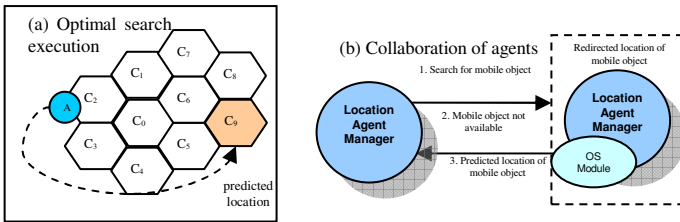


Fig. 4. Agents collaborate and tell the other agent of the possible location

```

INPUT: C All preprocessed data, i Mobile ID
OUTPUT: Predicted Location node, PNode
//preprocess the data
C = List all reci(t1(loc), t2(loc),...t6(loc))
//process the optimal search that has a UMP mining
PNode = OS(C)
//consult to parent agent
If PNode is Null Then GoTo ParentNode
//the agent knows and search in the node
Else Return Node Location and Search i

//inserted function of collaboration
Function AgentSearch (MobileID)
{
GoTo Node
Search all List mobileID in Database
If MobileID(List) = MobileID Then Return Node Location
Else PNode = OS(C)
Return PNode
}
    
```

Fig. 5. Optimal search algorithm with collaborative function

retrieval. It is done by communicating on the higher-level agents in the hierarchical structure of the location management.

Figure 4 illustrates the process of the optimal search method and the collaboration of agents. In the first process, Figure 4(a), the LAM executes the optimal search by mining the mobility patterns of previous mobile request. This results a the probable location of the mobile object. Figure 4(b) presents the collaboration function of the agents for the other agents to know the probable location of the mobile object. Figure 5 presents a pseudo-code of the optimal search and collaboration of agent.

4 Performance Analysis

The proposed system includes a search method using an optimal search algorithm and collaboration of agents. We assumed a hierarchy of location database appropriately placed at each node in a mesh network. To allow maximum flexibility in the design of the proposed location management, we considered hierarchies with a variable number of levels. The region is covered by databases which correspond to a unique physical address.

4.1 Hierarchical System Model Analysis

We compared the performance of hierarchical search, a nearest neighbor [13] and the proposed optimal search. The number of hops is determined by the level of the nodes and the search method is done from the lowest-level node up to the highest-level. In our simulation we have inputs of 15 mobile objects per nodes and considering the total number of database operation is equivalent to 0.01 of a second to process the mobile object as outputs. Equation 2 represents the simulation formula.

$$t_{sim} = \frac{1}{10} (n_l D_{op} + D_{op} (cn_{left}, cn_{right})) \quad (2)$$

The t_{sim} indicates the simulation time of the system model and l as the level of node n . The total database operation which indicated by D_{op} , is increasing as the level of the nodes increases because it has the data of every child nodes. The local database operation is added by $D_{op}(cn_{left}, cn_{right})$ which indicates the database cost of its left and right child nodes. The results of the simulation are presented in Figure 7.

5 Experimental Evaluation

The simulation used a database consisting of 22 nodes and 20 mobile ID in each were stored within the network. The simulation platforms used in the research were IBM compatible PC with Windows and Linux operating systems for the nodes, Borland Visibroker for implementing CORBA, Java for development of software agents, MySQL for the database storage of the system, and Weka Explorer to simulate the preprocessing and user mobility pattern and rules extraction. Figure 6 shows the interface of location agent manager implemented in CORBA Java.

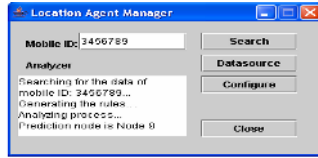


Fig. 6. Interface of location agent manager

5.1 Results of the Optimal Search Method

First we executed the OSM to mine the user mobility pattern in the mobile request data and produced the mobility rules of user A. User A has a 322 attributes in the mobile request database and we selected the first 6 hours of user A location. The preprocess method only selects the latest 100 tuples of data and ignoring the data with missing data values. The selection of many tuples will increase the time processing so that is why the location agent manager was configured with only limited records. After selecting, the Apriori configuration was set with the minimum support of 50 out of 100, which means that the possibilities of the pattern will most likely 50 percent and high patterns can be found in the data. The result produces 63 mobility patterns and 541 rules. These rules are filtered by the location agent manager according to the matching existing patterns to find the best predicted mobile location. The execution time of mining the pattern and rules of 100 tuples and filtering this rule is averaging at 1.15 seconds to execute. Table 1 presents the result of the rules generated.

Table 1. The user mobility rules generated

541 rules of user A mobility, showing only 5	Support	Confidence
1=N10 2=N1 4=N9 ==> 5=N2	0.69	0.99
1=N10 2=N1 4=N9 5=N2==> 6=N5	0.685	0.99
1=N10 2=N1 3=N6 66 ==> 4=N9 5=N2 6=N5	0.655	0.98
1=N10 2=N1 ==> 4=N9 5=N2	0.70	0.97
1=N10 ==> 4=N9	0.76	0.95

Secondly, the experiment used the hierarchical search method (HSM) of a mobile ID in the network, the nearest neighbor nodes (NN) method [13] and then the proposed optimal search methods (OSM). The LAM, knowing the mobility patterns of the user, predicts the next mobile location. The simulation result is shown in Figure 7. The graph shows the total time result of the three methods. It implies that an increase of search time on the next level of node is observed because of the incremental database operation from the child nodes in HSM while there is a minimized simulation time in OSM. The total time on performing until 6th level of the nodes are, 27.9 (OSM), 40 (NN) and 78.8 (HSM) seconds, respectively.

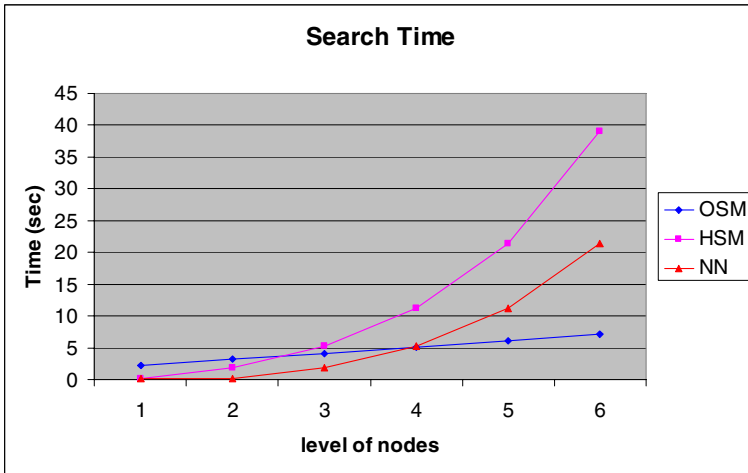


Fig. 7. Search time performance

6 Conclusions and Future Works

In this paper, we presented a collaboration of agents in the distributed location management to optimize the process of location management. A hierarchical scheme of searching and updating mobile objects was used and an optimal search method which locates the probable location of mobile object for retrieval was also presented. The OSM was done by the location agent manager which learns from the user mobility patterns and predicts the next location node of the mobile object. We used the Apriori algorithm in the OS module to mine the data. The result of the simulation and comparison with the other method implies that the technique using the proposed optimal method minimizes the searching time in the hierarchical scheme location management.

In the optimal search method, Apriori algorithm was only applied and other method of data mining may be applied in the module to test the efficiency of prediction. Also, too much random of the mobile user movements will make a lower confidence which implies that the prediction is not accurate and this will be the future study to solve these problem.

References

1. Jensen, C., Christensen, A., Pedersen, T., Pfoser, D., Saltenis, S. and Tryfona, N.: Location-Based Services: A Database Perspective. Proceedings of Scandinavian GIS, (2001)
2. Roth, J.: Flexible Positioning for Location-Based Services, IADIS International Journal on WWW/Internet, Vol. 1, No. 2, (2004) pp. 18-22
3. Jaric, P.: An Agent-Based Location System. Uppsala University, unpublished Master's Thesis in Computing Science 141, (1999)
4. Yavas, G., Katsaros, D., Ulusoy, O., and Manolopoulos, Y.: A Data Mining Approach for Location Prediction in Mobile Environments. Data & Knowledge Engineering 54, (2005) pp 121-146.

5. Krishnamachari, B., Gau, R., Wicker, S., and Haas, Z.: Optimal Sequential Paging in Cellular Networks. *Wireless Networks*, Vol. 10 , Issue 2, (March 2004) pp. 121 - 131
6. Liang, B., and Haas, J.: Predictive Distance-based Mobility Management for PCS networks. *Proceeding of the Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, Vol. 3, (1999) pp.1377-84
7. Wan, G. and Lin, E.: A Dynamic Paging Scheme for Wireless Communication Systems. *Proceeding of ACM/IEEE International Conference on Mobile Computing and Networking*, (1997) pp.195-203.
8. Zhang, J.: Location Management in Cellular Networks. *Handbook of Wireless Networks and Mobile Computing*, (2002) pp. 27-49
9. Pitoura, E., and Fudos, I.: Distributed Location Databases for Tracking Highly Mobile Objects. *The Computer Journal*, Vol. 44, No.2, 2001.
10. Lee, K., Lee, H., Jha, S., and Bulusu, N.: Adaptive, Distributed Location Management in Mobile, Wireless Networks. *Proceedings of the 2004 IEEE International Conference on Communications*, (2004), pp. 4077-4081
11. Das, S. and Sen, S.: Adaptive Location Prediction Strategies Based on a Hierarchical Network Model in a Cellular Mobile Environment. *The Computer Journal*, Vol. 42, No. 6, (1999)
12. Mateo, R. M., Lee, J. W. and Kwon, O.: Hierarchical Structured Multi-agent for Distributed Databases in Location Based Services. *The Journal of Information Systems*, Vol. 14, Special Issue (December 2005) pp. 17-22
13. Gerardo, B. D., Lee, J. W. and Joo, S.: The HCARD Model using an Agent for Knowledge Discovery. *The Journal of Information Systems*, Vol. 14, Special Issue (December 2005) pp. 53-58.
14. Want, R., Hopper, A., Falcao, V. and Gibbons, J.: The Active Badge Location System. *ACM Transactions on Information Systems*, Vol. 10 , Issue 1 (January 1992), pp. 91-102

Design of H.264/AVC-Based Software Decoder for Mobile Phone*

Hyung-Su Jeon, Hye-Min Noh, Cheol-Jung Yoo, and Ok-Bae Chang

Dept. of Computer Science, Chonbuk National University, 664-14 1ga, Duckjin-Dong,
Duckjin-Gu, Jeonju, Jeonbuk, South Korea
{hsjeon, hmno, cjyoo, okjang}@chonbuk.ac.kr

Abstract. Investigations of the video service technology used for mobile phone have been actively performed recently. The efforts on the application of all the technologies available for commercialization of the framework of wire-internet into mobile environment have been also researched, thanks to the speedy development of a mobile phone and wireless network platform. The video service related technology based on the mobile phone has been implemented and operated on the basis of hardware.

However, Current service mode, which is based on hardware, has its own disadvantage of not coping flexibly with various changes in the control structure of the new video codec algorithm and video data communication. Accordingly it is necessary to develop a decoder, which can deal with video service technology embodied in the form of hardware into that of software. This requires us to develop a new video player. In addition, such a decoder can achieve an immediate response due to further technology development including video data transmission and traffic control without consuming additional resources, in comparison with a hardware decoder chip technology. It can greatly contribute to an improvement in the manufacturing cost arising from incorporating an additional chip into mobile phones, as well as the reduction of resources through recycling. This paper designed a H.264/AVC video software-based decoder that uses component base design at the WIPI platform.

1 Introduction

To realize mobile video service, it is necessary to achieve a high efficient video encoding method, a low power means of consumption for mobile phones and a decoding method with a low complexity. To address these requirements suitable for a mobile phone environment and multiple clients, the next generation video compression technology, H.264/AVC has emerged.

The bit rate and a picture size supported by H.264/AVC have a wide application range and have the capability of extending a very low bit rate, a low frame rate and a wide range of video compression. H.264/AVC video compressing technology is being rated higher in performance than previous MPEG-4. In Korea, H.264 has been not

* This work was supported by grant No. R01-2004-000-10730-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

only designated as S-DMB standard video decoder of a satellite broadcasting service, but also has been considered as one of the core multimedia technologies in mobile industry by being designated as T-DMB standard video decoder. Currently several mobile communication service providers have been offering a part of streaming services with H.264/AVC so that it has become one of the vital technologies in the mobile communication industry.

H.264/AVC has been applied in the hardware of mobile phone and this hardware-based decoder has presented the disadvantage of requiring a great deal of investment, and causing the waste of resources because it calls for the replacement into a new mobile phone with the development of compression technology. Accordingly it is necessary to develop the decoder based on software, which can be applied to mobile phones as being required for the realization in the form of software.

This paper designed a software-based decoder for mobile phone in the component structure by applying H.264/AVC video compression technology on the basis of the WIPI platform as the next generation mobile standard.

2 WIPI Platform and H.264/AVC

2.1 WIPI Platform

A platform of mobile phone in each company uses an independent one as an application development platform for a mobile phone. This means that each company develops its own platform or adopts a unique one suitable for its own application. However, such a non-integrated platform does not provide compatibilities among different mobile phones from each service provider that mobile phone subscribers are not allowed to use the contents they would like to use with their mobile phones. CP(contents provider) has to develop the contents acceptable to various platforms even for the same contents, which causes to spend more time and cost in doing so. Table 1 shows a comparison among the mobile phone platforms prevailed in recent days[1].

Table 1. Comparison among Mobile Phone Platforms

Platform	XVM (SK-VM)	KVM (JAVA Station)	BREW
Developer	XCE	SUN	Qualcomm
Language	JAVA	JAVA	C
Type	Script	Byte Code	Binary
Features	Sun MIDP Sole Realization	kitty Hawk Porting	Supporting runtime environment
Advantage	Service available according to the content type	Duplex page, Complete portability	Applicable to the world-wide using CDMA chip, Easy to develop contents

Vigorous discussions about the standard platform, which does not cause any confusion to CPs and prevents mobile phone manufacturers from spending excessive engineering resources, have been brought forth in Korea, which finally led to the development of the WIPI platform.

The WIPI platform defines the specification of a mobile standard platform enabling the environment to perform an application program while being installed on a mobile communication terminal. To the application program developer, it ensures the contents compatibilities among platforms, easy-to-porting to mobile phone developers and to the subscribers, extensive and sufficient services. A standard platform of a mobile phone has in general the following structure outlined in Fig. 1. In the basic S/W of a mobile phone includes communication function and every device driver[2].

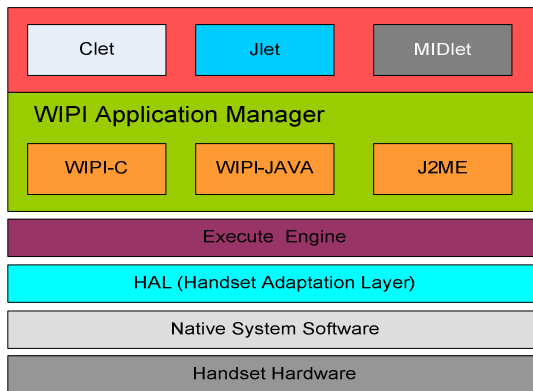


Fig. 1. A Conceptual Structure of A Standard Mobile Phone Platform

The following describes the technical features of the WIPI platform.

- Multi language support of C, C++, and Java
- Quick execution speed
- Execution of multiple applications
- A solid security model
- Qualified as an application software

As seen in Fig. 2, the WIPI platform defines only a minimum API set to maintain the compatibility of the contents. That is, it defines the specification of an interface only for a smooth contents development. For example, the object to be finally downloaded on a mobile phone requires the form of a machine code as one of the requirements from the service provider. As the principles are adopted only for this requirement, the developers attempting to realize a standard platform can actually develop various solutions within the scope of ensuring the compatibility of the contents. Also a general requirement prescribed in the standard specification requires both a platform and an application to be designed independently from a hardware, making it possible to execute and port easily regardless OS which is used by the hardware of a mobile phone including CPU, LCD and memory or a mobile phone.

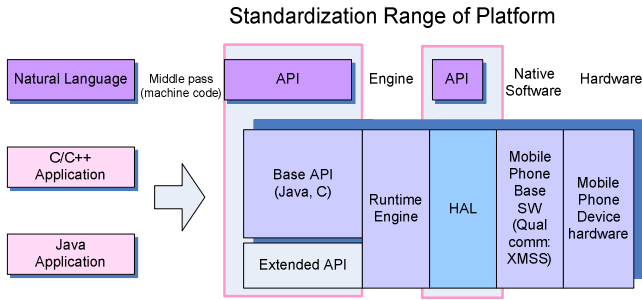


Fig. 2. A Standardization Range of the WIPI Platform

2.2 H.264/AVC

H.264/AVC is a standard for the coded expression of visual information. This standard specifies the syntax of coded bitstream, the semantics of these syntax elements and the process by which the syntax elements may be decoded to produce visual information[3,4,5].

2.2.1 A Comparison Between MPEG-4 Visual and H.264/AVC

H.264/AVC is an algorithm based on an effective compression of a video frame. Accordingly the standard focuses on the production of a popular video compressed application program with compression and transmission efficiency. Table 2 summarizes the main difference compared MPEG-4/Visual with a similar standard attempting to indicate the feature of an H.264/AVC's algorithm by which we can come to understand the intention of how to compress and understand the applicability pursued by an H.264/AVC, even though a full comparison is not represented.

Table 2. Comparison of difference between MPEG-4 Visual and H.264/AVC

Comparison	MPEG-4 Visual	H.264/AVC
Number of profiles	19	3
Compression efficiency	Medium	High
Support for video streaming	Scalable coding	Switching slices
Motion compensation minimum block size	8x8	4x4
Motion vector accuracy	1/2 or 1/4 pixel	1/4 pixel
Built-in deblocking filter	No	Yes

2.2.2 Profiles of H.264/AVC

H.264/AVC defines a set of three profiles(Main, Extended and Baseline), each supporting a particular set of coding functions and each specifying what is required of

an encoder or decoder that complies with the profile. Table 3 is a list showing the applications to be selected in a video coding technology, the major requirements of each area and the profiles of an H.264/AVC suitable for the areas. The software decoder designed in this paper selected a baseline profile proposed as being suitable for a mobile video in the standard specification of H.264/AVC[7,8,9].

Table 3. Applicable Areas and Requirements

Application	Requirements	H.264/AVC Profile
Broadcast television	Coding efficiency, reliability, interlace, low-complexity decoder	Main
Streaming video	Coding efficiency, reliability, scalability	Extended
Video storage	Coding efficiency, interlace, Low-complexity decoder	Main
Videoconferencing	Coding efficiency, reliability, low latency, low-complexity encoder and decoder	Baseline
Mobile video	Coding efficiency, reliability, low latency, low-complexity encoder and decoder, low power consumption	Baseline
Studio distribution	Lossless or near-lossless, Interlace, effective transcoding	Main

Mobile phone’s environment is far behind in performance than that of an ordinary computer system. A long period of time is required to load data and it presents a seam in audio code due to low performance and memory capacity. As it exhibits a relatively low transmission speed in comparison to that of a network and the internet, a focus should be placed on the possible reduction of data capacity and the optimization of a buffering program for video to be smoothly played. Accordingly as presented in Table 3, H.264/AVC with higher compression efficiency can be regarded as being better than any other for mobile environment.

2.2.3 Structure of H.264/AVC Decoder

Fig. 3 shows the structure of H.264/AVC decoder. H.264/AVC decoder[3,4] receives a compressed bitstream from NAL(Network Abstraction Layer) and entropy decodes the data elements to produces a set of quantized coefficients X. These are scales and inverse transformed to give D’n. Using the header information decoded from the bitstream, the decoder creates a prediction block PRED, identical to the original prediction PRED formed in the encoder. PRED is added to D’n to produce uF’n which is filtered to create each decoded block F’n[5,10].

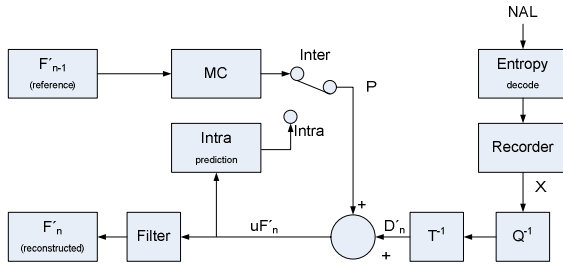


Fig. 3. Structure of H.264/AVC Decoder

3 Design of Software Decoder

Wireless mobile platform, the type of encoding data which decodes the encoded data, the means of network transmission and the selection and level of an appropriate H.264/AVC profile should be considered to design H.264/AVC software decoder.

Once designing H.264/AVC in consideration of the above criteria, the center module broadly consists of the following three components of a network manager: an event process and a decoder of which a decoder has a crucial role to process the encoded data to be able to receive in the form of stream for the input of video into a decoder from a network manager. It plays the role of decoding the encoded data for which a decoder receives the output and of showing it to a mobile phone according to an external key event.

An event processes, as the device of transmitting the event coming from an event process module of video player UI to a decoder, extends an interface role between video codec and UI[5]. This is because a decoder does not bring any effect on an external UI even though being upgraded in the future by being processed through the interface of an event process of a decoder module without processed directly in the video player. This makes it possible to achieve performance improvement and easy maintenance by modifying the design in this way, even though there may be possible changes in UI of video player.

The standard specification of H.264/AVC does not define how to transmit a NAL unit. This means that any type can realize the method of transmitting a NAL unit. A network manager plays a role in transmitting a NAL unit and is designed as an independent component from a decoder.

Baseline profile is the one suitable for mobile video suggested in the H.264/AVC standard as precisely mentioned. This paper applied the baseline profile for designing a decoder and placed several steps appropriate to mobile phone for the various levels of parameters such as sample processing rate, picture size, compression bit rate, and memory requirement.

Summarizing the software decoder designed in this paper reveals that it is separated into three functions of a network manager. These three functions consist of an event process and a decoder for H.264/AVC Baseline profile and designed in the type of a component. A network manager was designed to control video data according to the status of network after streaming processing. With the video data decoded with RTP protocol, an event process was designed to process all the video events including

play, stop and pause, which come from video player. The center module of a decoder, in consideration of the performance and features of mobile phone, was designed after being reconstructed in the modules of an error concealment processing, a multi-slice group supporting, a macro block processing and CABAC entropy coding excluding loop one of video filter. Fig. 4 indicates the structure of H.264/AVC software decoder.

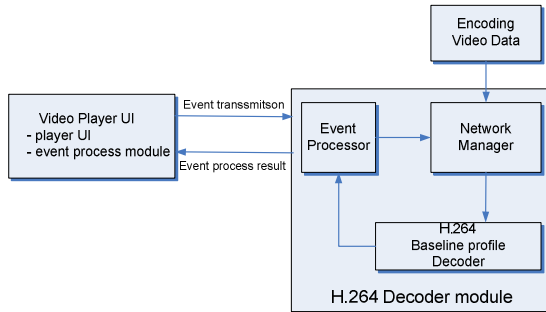


Fig. 4. Structure of H.264/AVC Software Decoder

Once designed, the decoder is set to decode a video by executing a video player at the WIPI platform. The following figures highlight the skeleton codes of a video player, a network manager component, a decoder main and event process class and the interfaces, which have been defined above.

```
import org.kwis.msp.lcdui.*;
import org.kwis.msp.lwc.*;
import kr.codec.ldecode.*;
public class MediaPlayer extends Jlet {
    protected void startApp(String args[]) {
        Display dis = Display.getDefaultDisplay();
        Decoder dec = Decoder.getInit();
        dis.pushCard(new Card()) {
            // set up initial display
        }
        // video playing with transmission manager and
        decoder
    }
    protected void pauseApp() {
    }
    protected void resumeApp() {
    }
    protected void destroyApp(boolean b) {
    }
}
```

Fig. 5. Video Player Component

```

import org.kwis.msp.lcdui.*;
import org.kwis.msp.lwc.*;
import kr.codec.ldecode.*;

public class NetworkMngr extends DecoderCommon {
    // Realization of network manager
}

```

Fig. 6. Network Manager Component

```

import org.kwis.msp.lcdui.*;
import org.kwis.msp.lwc.*;
import kr.codec.ldecode.*;

public class DecoderMain extends DecoderCommon {
    // Realization of decoder main
}

```

Fig. 7. Decoder Main Class

```

import org.kwis.msp.lcdui.*;
import org.kwis.msp.lwc.*;
import kr.codec.ldecode.*;

public class MessageCtrl extends DecoderCommon {
    // Realization of event process
}

```

Fig. 8. Event Process Class

```

public interface DecoderCommon {
    // Defines a set of constants used for decoder
}

```

Fig. 9. Common Decoder Constant Interface

The decoder designed in this research has an additional component for detailed function except the 5 classes and the interface mentioned above of which a module to support slice group, a macro block module, a CABAC entropy coding module, a file processing module and a parameter set module are ones of the crucial modules.

The overall structure of a system consists of a network process, a video player, a software decoder and a data structure process. The critical function of a software decoder, as seen in Fig. 10, is consisted of 4 parts; a macro block process, a parameter set process, a buffer process, and an error concealment process. A loop filter considered in a decoder is excluded from this construction on the basis of performance.

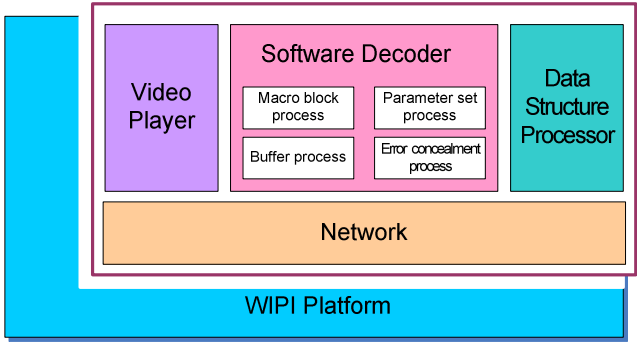


Fig. 10. Overall Structure of Software Decoder System

A macro block process decodes the encoded macro block, create the originally predicted block and plays the role of decoding video, and a parameter set process interprets the parameter applied to overall video images and is used for restoring video images. A buffer process stores a streaming data transmitted through the network at a temporary place and play it, and an error concealment process is used for restoring the phenomenon which appears as a still image as video data transmitted through the network can not be decoded with a big difference between the current frame and the old one, due to errors like intervention.

4 Conclusion and Further Suggestions

A software decoder based on a hardware chip was tested for use recently in mobile phones to provide a video service. Use of this type of service has been frequent such that the same quality service for the same contents could not be guaranteed across the board for various mobile phone manufacturers, and the subsequent service methods offered by the service providers. Accordingly it is necessary to pinpoint the method of receiving a video service without depending on the type of mobile phone at the wireless network platform. It is, therefore, highly necessary to develop a software-based decoder casting off a previous hardware-based decoder. H.264/AVC is an appropriate type as making it possible to play a video in a smooth manner even under a low transmission speed by reducing data volume through increasing compression efficiency. The WIPI has been suggested as a standard platform over several consultations as there has been a difficulty in developing the application for platform variations prevailed in domestic market.

This paper designed the software-based decoder for mobile phones on the basis of the WIPI platform. The software decoder was classified into a network manager, an event process and H.264/AVC decoder for baseline profile, and was designed in the form of a component.

In order to realize the decoder suggested here in this paper, the following recommendations should be considered together with the decoder itself; how to achieve an effective transmission in a video communication system under mobile environment, a low power problem of mobile phone and calculation capability. Further studies should be deployed to ensure the achievement of an efficient switching among several coded streams by making H.264/AVC furnish SI and SP-slice in order to solve these problems in the future. Also to optimize the speed of a mobile phone, studies on the set-up of a library which controls a process function at a low level by utilizing directly the command of specific processor incorporated in mobile phone including ARM processor must be conducted in the future.

References

1. Si-woo Byun and Sook-eun Byun, "A study on WIPI Platform for Efficient Mobile Business," KSI, Vol. 4, No. 2, pp. 79-93, 2003.
2. Seok-hee Bae "Trends in Mobile Platform Standardization and Future Direction for its Development," TTA Journal, No. 82, pp. 20-30, 2002.
3. ISO/IEC 14496-10 and ITU-T REC, H.264 Advanced Video Coding, 2003.
4. GARY J. Sullivan "Video Compression - Form Concepts to the H.264/AVC Standard," Proc. of the IEEE, Dec. 2004.
5. A. Hallapuro, M. Karczewicz, and H. Malvar, "Low Complexity Transform and Quantization - Part I : Basic Implementation," JVT document JVT-B038, Geneva, Feb. 2002.
6. Iain E. G. Richardson, H.264 and MPEG-4 Video Compression, John Wiley&Sons, 2003.
7. M. D. Walker, M. Nilsson, T. Jebb, and R. Turnbull, "Mobile Video-Streaming," BT Technology Journal, Vol. 21, No. 3, pp. 192-202, Jul. 2003.
8. T. Stockhammer, M. M. Hannuksela, and T. Wiegand, "H.264/AVC in Wireless Environments," IEEE Trans. CSVT, pp. 657-673, Jul. 2003.
9. C. Kim and J. N. Hwang, "Fast and Automatic Video Object Segmentation and Tracking for Content-Based Applications," IEEE Trans. CSVT, pp. 122-129, Feb. 2002.
10. S.W. Golomb, "Run-length encoding," IEEE Trans. on Information Theory, IT-12, pp. 399-401, 1966.

Transforming a Legacy System into Components

Haeng-Kon Kim¹ and Youn-Ky Chung²

¹Department of Computer Information & Communication Engineering,
Catholic University of Daegu, South Korea
hangkon@cu.ac.kr

²Department of Computer Engineering, Kyung Il University,
Kyungsan, Daegu, 712-701, Korea
ykchung@kiu.ac.kr

Abstract. Most legacy systems are being pressured to continuously respond to changing requirements, but it is impossible almost to cope with these requests effectively. Because many legacy systems have suffered from lack of standardization and openness, difficulty of change, and absence of distributed architecture. Especially, according as legacy system has been deteriorating from an architectural point of view over the years, we must continually maintain these legacy systems at high cost for applying new technologies and extending their business requirements. For the purposes of transforming a legacy system into component system, we need systematic methodologies and concrete guidelines. Through these, we can share information at different levels of abstraction ranging from code to software architecture, and construct the component system with better component-based architecture.

To achieve these goals, we have built upon the L2CBD (Legacy to Component Based Development) methodology providing reengineering process including concrete procedures, product-works, guidelines and considerations. We can transform legacy systems into new component system with improved software architecture by adapting L2CBD.

1 Introduction

A legacy system is an application that was developed on older technology and is past its prime use, but still play an important part in current businesses[1]. Therefore, this legacy system is being viewed as an asset that represents an investment that grows in value rather than a liability whose value depreciated over the times[2]. But these legacy software systems have suffered from lack of standardization and openness, difficulty of change, and absence of distributed architecture. Instead of continually maintaining these legacy systems at high cost for applying new technologies and extending their business requirements, reengineering them to new systems with good design and architecture can improve their understandability, reusability and maintainability. If we want to accommodate future changes in the software, we must have the well-defined architecture that is of the utmost importance with respect to flexibility and maintainability of software. Therefore, the main issues in current reengineering approaches are how to

integrate legacy system with emerging technologies for reflecting incremented requirements, reuse the legacy assets for producing software system to be required in future, and improve the availability and quality of legacy system.

In this paper, we propose the L2CBD as reengineering methodology for transforming legacy system into component system. It features an architecture-based approach to deriving new application structure, a reverse engineering technique for extracting architectural information from exist code and business domain knowledge, an approach to component system generation that system's architectural components can be reused in the evolved version, and a specific technique for dealing with the difficulties that arise when extracting components from legacy system and deciding transformation strategies and process. Because L2CBD provides the concrete reengineering steps including definite working procedure, relationship among work-products, practical transformation guidelines, we can construct a component system with better architecture from legacy system by adapting it.

2 Related Works

A Reengineering is the systematic transformation of an existing system into a new form to realize quality improvements in operations, system capability, functionality, performance or evolvable maintenance at a lower cost, schedule, or risk to customs[4].

There is CORUMII (Common Object-based Reengineering Unified Model II)[5] defined in SEI CMU as a reengineering methodology to be referred most widely. CORUM model is a reengineering tool interoperation to include software architecture concepts and tools. The extended framework - called CORUMII- is organized around the metaphor of a "horseshoe", where the left-hand side of the horseshoe consists of fact extraction from an existing system, the right hand side consists of development activities, and the bridge between the sides consists of a set of transformations from the old to the new. As another reengineering methodology, there is MORALE(Mission ORiented Architecture Legacy Evolution)[6] that developed in Georgia Institute of Technology. It features an inquiry-based approach to eliciting change requirements, a reverse engineering technique for extracting architectural information from exist code, an approach to impact assessment that determines the extent to which the existing system's architectural components can be reused in the evolved version, and a specific technique for dealing with the difficulties that arise when evolving user interface.

3 Transforming a Legacy into Components

Fig. 1. is shown the overall concept model of L2CBD. L2CBD is not sequential process like existing methodologies, but it is parallel and customizable process adjusting by customer's requirements. So, it is architecture-oriented reengineering process based on CBD and it supports evolutionary extension, assembly of components and customization of process toward target system.

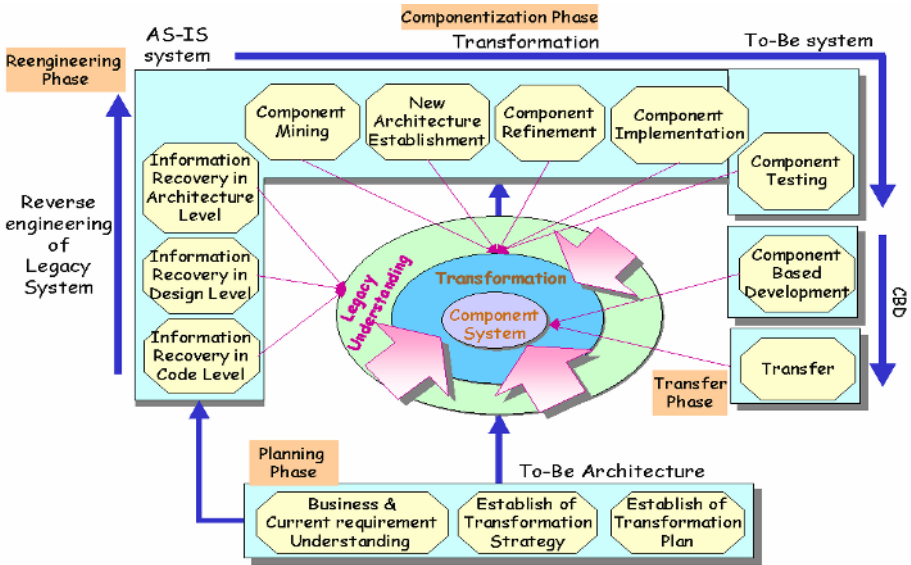


Fig. 1. Overall concept model of L2CBD

3.1 Customization Process

Phase and evaluation of L2CBD is shown as Fig. 2. It is not sequential process, but it is selectable, iterative and customizable process. The other words, L2CBD supports continuous extension, assembly and customization of components based on target architecture.

Accordingly, reengineers can acquire various procedures for communication among project stakeholders, and the best guidelines tailored to their specific environment. We may perform the componentization phase after finishing reverse engineering phase based on reengineering strategies and process, or first, we directly perform componentization phase and we may acquire necessary information by achieving the reverse engineering phase whenever we need. Also, component phase and transfer phase are selected or skipped or iterated by necessary[9].

3.2 Overall Configuration

Metamodel of L2CBD is like a Fig. 3. A reengineering project is realized through reengineering process that is consisted in several phases. A phase is a unit separated logically in reengineering process and it includes several activities that are a set of systematical tasks grouped for specific goal. Also, task has procedures representing more detail techniques and order, guidelines representing essential constraints and requirements, and work products produced through performing each work[10]. Users of L2CBD can compose the their own reengineering process by selecting and repeating these phase, activities and tasks.

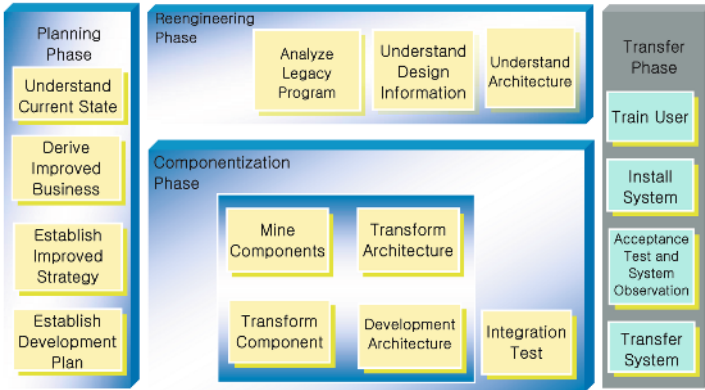


Fig. 2. Phase and evaluation of L2CBD

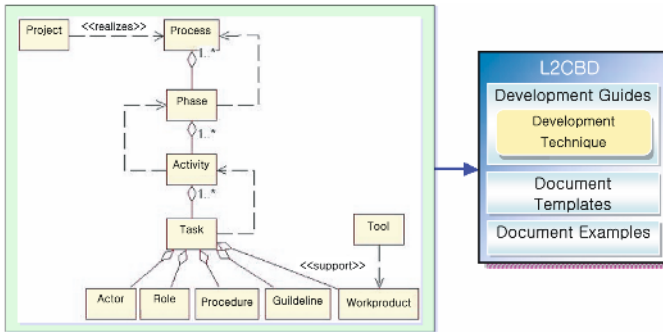


Fig. 3. Metamodel of L2CBD

3.3 Planning

In this phase, we capture some problems contained in current system and understand effects to be expected through reengineering. Also, decide the objectives and extent of reengineering project, and propose the componentization strategy, improvement directions and evolution process.

Fig. 4 presents tasks of planning phase. Each activity is addressed more detail as following.

(1) Understand Current State activity

To cope with market change, increasing customer requirements, and introduction of new technologies, we should analyze and understand the environment of legacy system in aspect of business, technical and operational view. To do this, we derive the internal issues and main problems which organization has faced; analyze the organization’s strength and weakness; understand the functionalities of business and legacy subsystem; and grasp the system environment of legacy system and maintenance status.

(2) Derive Improve Business Model activity

This activity includes 4 tasks. In the first task, grasp the initial requirements for future.

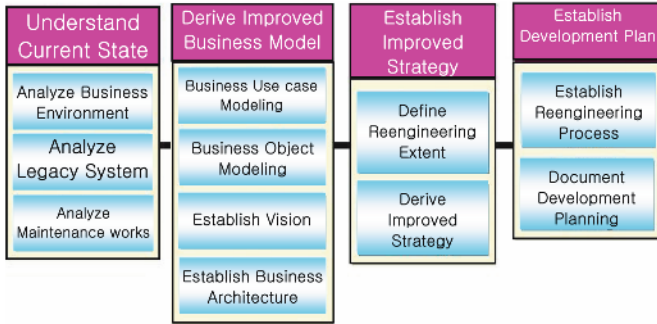


Fig. 4. Tasks of Planning Phase

3.4 Reverse Engineering Phase

This phase is performed to obtain understanding information of legacy system for maintaining them and adapting them into new environments. To do this, grasp the main functionalities and structure of legacy system. And, understand the static and dynamic information of legacy system by analyzing source codes and modeling design information. Also, take a preparation tasks for componentization through understanding and abstracting the relationships between elements of legacy system.

(1) Analyze legacy Program activity

This activity includes general tasks happened in reverse engineering process. Therefore, we perform an information analysis of program unit and system unit by eliminating non-used codes, structuring the codes, and analyzing a flow and control of program. In this activity, we output a re-structured legacy code, relation table of variables, structure chart, call graph, control graph, and screen flow graph of program.

(2) Understand Design Information activity

The goal of this activity is acquiring the base information to understand and construct the architecture of target system by extracting the analysis and design information of legacy system. If documentations of legacy system are not enough and these documents are not corresponded to extracted information from legacy codes, this activity is specially needed. We complete Entity Relation table by extracting entities, properties by entity, and relationships among entities, and diagrammatizing them. Also, generate database scheme of legacy system, and use case mapping table, which summarize mapping relations between business use cases and application use cases of legacy system.

(3) Understand Architecture activity

Through this activity, we can completed understand of legacy system by extracting architectures of various aspects. To do this, write a structural architecture by identifying the subsystems consisted in legacy system, and grasping the hierarchy relationship

and dependency among them. And write behavior architecture through identifying call relationship among subsystems. Lastly, generate technical architecture represented what techniques are applied to develop hardware devices and subsystems be placed in these devices.

3.5 Component Phase

The previous two phases are focused on extracting the information from legacy system and abstracting this information for understanding of legacy system in various aspects. But main activities of this phase are transforming legacy information including all outputs as likes source codes, design modeling et al., to new forms required in target environment.

Our goal is transforming non-component legacy systems into component-based system. Therefore, in this phase, we establish new software, component and system architecture based on analyzed information in Planning phases. And extract candidate components and transform them to target component to be adapted in new target architecture.

Fig. 5 presents all activities and tasks consisted in Component phase.

(1) Mine Component activity

This activity has the first tasks to transform legacy system into target system with new architecture. In this activity, we divide legacy system into independent functional unit, and these each is mapped into a component candidate. Component candidates are a group of system elements with high dependency. And based on business use case generated in Planning phase, we can identify elements with independent business functionality as component candidates. Also, componentization strategy of each component candidate extracted is redefined.

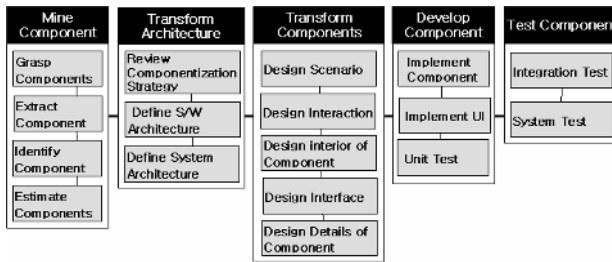


Fig. 5. Tasks of Component Phase

(2) Transform Architecture activity

In this activity, Componentization technique(that is Transformation or Wrapping) is conformed by comparing and analyzing the componentization strategies each defined in Planning phase and Mine Component activity. And, decide new architecture style of target system through analyzing the requirements from point functionality and quality of view. To do this, define software architecture, design a component and technical architecture, and generate system architecture integrated all architecture viewpoints.

(3) Transform Component activity

This activity has 5 tasks as following: First, In Design scenario tasks, write each specification about business use cases extracted Planning and application use cases of legacy system derived Reverse Engineering phase. Next we represent what interaction relationships among entity of each use case are happened, through sequence diagram and collaboration diagram, in Design Interaction task. Sequentially, design the internal structure of classes composed a component through Design Interior Component task, and identifies component interfaces and operations included in them based on dynamic message flow of internal classes of component in Design Interface task. In Design details of Component task, adapt design information designed in previous tasks to specific component platforms by defining package and EJB mapping relations, and design persistence, transaction, security and deployment.

(4) Develop Component activity

This activity has real development task of target components. To do this, complete programs of components and classes and eliminate syntax errors of components. And implement UI screens of target system and integrate them into corresponded components. Also, perform unit test of components and classes focused on all lines of source codes.

(5) Test Component activity

Through this activity, we test that communications of component interfaces among related components are correct and target system constructed by integration of each component performs the same functions to legacy system. Also, assure that software architecture is robust and all business rules are implemented well. Beside, check that various stakeholders are satisfied by target system in aspects of functional, technical and quality requirements.

4 Conclusion

The most legacy software systems have suffered from lack of standardization and openness, difficulty of change, and absence of distributed architecture. Instead of continually maintaining these legacy systems at high cost for applying new technologies and extending their business requirements, reengineering them to new systems with good design and architecture can improve their understandability, reusability and maintainability. Especially, if we want to accommodate future changes in the software, we must have improvement architecture that is of the utmost importance with respect to flexibility and maintainability.

In this paper, we propose the L2CBD as reengineering methodology for transforming into component system. It features an architecture-based approach to deriving new application structure, a reverse engineering technique for extracting architectural information from exist code and business domain knowledge, an approach to component system generation that system's architectural components can be reused in the evolved version, and a specific technique for dealing with the difficulties that arise when extracting components from legacy system and deciding transformation strategies and process. Because L2CBD provide the concrete reengineering steps including definite working procedure, relationship among work-products, practical

transformation guidelines, we can construct a component system with better architecture from legacy system by adapting it.

References

1. Dolly M, Neumann, "Evolution Process for Legacy System Transformation", IEEE Technical Applications Conference, Washington, November, 1996, pp57-62
2. Nelson Weiderman, Dennis Smith, Scott Tilley, "Approaches to Legacy System Evolution", CMU/SEI-97-TR-014, 1997
3. William Ulrich, Legacy Systems : "Transformation Strategies", Prentice Hall, 2002
4. SEI Reengineering Center "Perspectives on Legacy System Reengineering, 1995
5. Rick Kazman, Steven G. Woods, S. Jeromy Carriere, "Requirements for Integrating Software Architecture and Reengineering Models: CORUM II", Fifth Working Conference on Reverse Engineering, Honolulu, Hawaii, Oct 1998, pp: 154-163
6. Abowd G. Goel A. Jerding D.F., McCracken M., Moore M., Murdock J.W., Potts C., Rugaber S., Wills L., "MORALE. Mission ORiented Architectural Legacy Evolution" International Conference on Software Maintenance, Bari, ITALY, October, 1997, pp150 –159
7. Jochen Seemann, Jürgen Wolff von Gudenberg, "Pattern-Based Design Recovery of Java Software", *Communications of the ACM*, Vol. 38, No. 10, pp65-74, October 1995.
8. Jung-Eun Cha, et al., "Reengineering Process for Componentization of Legacy System", *Journal of the Korea Society of System Integration*, Vol.2, No.1, pp 111 – 122, May, 2003
9. Jung-Eun Cha, et al., "Establishment of Strategies and Processes for Reengineering of Legacy System", proceedings of the 20th KIPS Fall Conference, Vol.10, No.2, Nov. 2003
10. Jung-Eun Cha, et al., "Definition of Metamodel for Reengineering Methodology of Legacy System", proceedings of the 5th KCSE Conference, Vol.5, No.1, Feb. 2003

Pseudorandom Number Generator Using Optimal Normal Basis

Injoo Jang and Hyeong Seon Yoo

School of Computer Science and Engineering, Inha University,
Incheon, 402-751, Korea
hsyoo@inha.ac.kr

Abstract. This paper proposes a simple pseudorandom number generator [PRNG] by using optimal normal basis. It is well known that the squaring and multiplication in finite field with optimal normal basis is very fast and the basis can be transformed to a canonical form. The suggested PRNG algorithm combines typical multiplications and exclusive-or bit operations, both operations can be easily implemented. It is shown that the algorithm passes all terms of the Diehard and the ENT tests for long sequences. This algorithm can be applied in various applications such as financial cryptography.

1 Introduction

There has been an increasing attention in the design of PRNG in various fields, statistical simulations, lottery, and cryptography. The application areas require different random properties, but it is required that PRNG should pass the standard statistical properties [1]. Linear congruence and recurrence equations are commonly adapted PRNG types; they often use modular arithmetic [2, 3, 4].

As big numbers are often required for random sequences, and it is common to employ one of two representation types such as polynomial basis and the optimal normal basis. The optimal normal basis eq.(1) has an equivalent shifted form of the canonical basis, eq. (2) [5, 6, 7].

$$M = \{\beta, \beta^2, \beta^{2^2}, \dots, \beta^{2^{n-1}}\} \text{ for some } \beta \in GF(2^n) \quad (1)$$
$$N = \{\gamma + \gamma^{-1}, \gamma^2 + \gamma^{-2}, \gamma^3 + \gamma^{-3}, \dots, \gamma^n + \gamma^{-n}\} \quad (2)$$
$$= \{\beta_1, \beta_2, \beta_3, \dots, \beta_n\}$$

The optimal normal basis can be very effective for squaring operations and the shifted form of the canonical basis can be easily adopted for multiplication process. The multiplication has one interesting property for randomness since the multiplied result is randomly distributed on the entire field coefficients. In this paper we are suggesting an PRNG with the optimal normal basis. The algorithm adapts the typical multiplication and exclusive-or operation [XOR] for giving randomness. Our idea is that if the multiplication of optimal normal basis is implemented in the algorithm, we could get randomness. And the XOR in the algorithm might be strength the random property.

2 A PRNG with Optimal Normal Basis

The commonly referred sequential PRNG is as eq. (3), [1]. The randomness of the sequence X_i will differ by the coefficients A, c .

$$X_{i+1} = (AX_i + c) \bmod m \quad (3)$$

The element multiplication in optimal normal basis is different from the polynomial types, the optimal normal basis could be a good factor for PRNG. We suggest two comparable PRNG counterparts in optimal normal basis type, eq. (4) and (5).

$$X_{i+1} = A \bullet (X_i \oplus B), \quad i \geq 1 \quad (4)$$

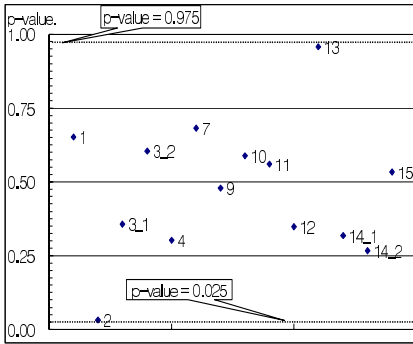
$$X_{i+1} = A \bullet X_i \oplus B, \quad i \geq 1 \quad (5)$$

All field values are written in the canonical type with 10 ~12 words of 32-bit to give long random number requirements. A, B, X_0 are random seeds which can be generated by a PRNG. Operation \bullet is the typical multiplication in optimal normal basis and \oplus is XOR in binary representation. The XOR and the multiplication costs can be reduced if we employ small number of words. Considering the efficient implementation, one of two random seeds A or B could be as small as one word. The multiplication process with optimal normal basis can be fast since there is no requirement for modular operation, which is proved very effective [6, 7]. The requirement for storage is minimal in the sense that we have only one random sequence.

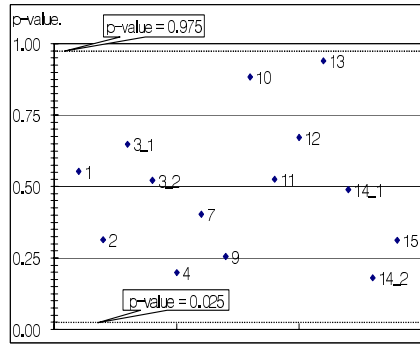
3 Testing Randomness of the Scheme

Usefulness of the suggested schemes should be proved by a standard test method, and there are a lot of research papers and standard methods on the randomness testing for a PRNG [8, 9, 10, 11, 12]. Diehard could be a test suite for randomness of eq. (4) and (5). It tests 15 items which include all aspects of randomness, and is tested and used in many applications for years [11]. So if any new scheme for PRNG passes all test items in Diehard, it is proved that the scheme is really pseudorandom in every aspect.

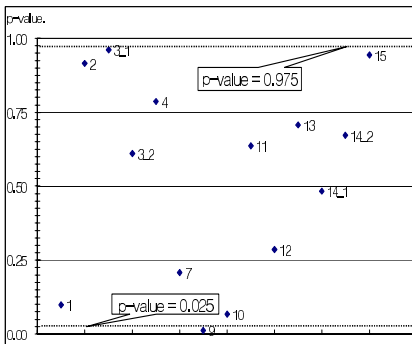
The Implementation of the schemes described here was done in C/C++ and run on a Pentium machine with CPU 2.00GHz, 512MB RAM. Data size of random sequence is chosen as 350 bit by considering optimal normal bases. Fig. 1 shows that all test results of eq. (4) are in acceptable range between 0.025 and 0.975, for different data sizes of random seeds. Random seed sizes are changed from full to one word. All combinations of different data sizes combinations pass the Diehard and the sequences are perfectly random. It passes even for both seeds are 1 word. Table 1 is the summary of Diehard test items and the nomenclature of data size.



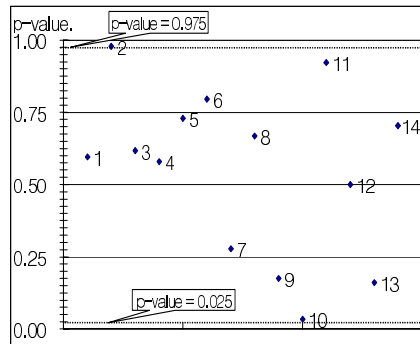
(f,1)



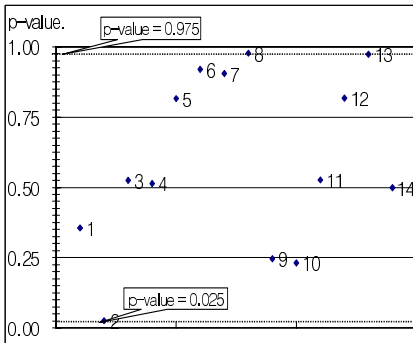
(1,f)



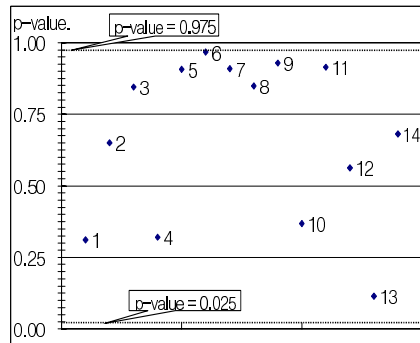
(2,2)



(1,2)



(2,1)



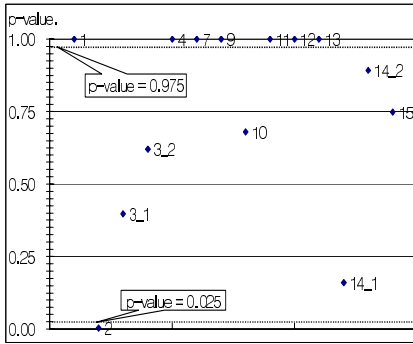
(1,1)

Fig. 1. Diehard results for data field sizes of random seed, eq. (4)

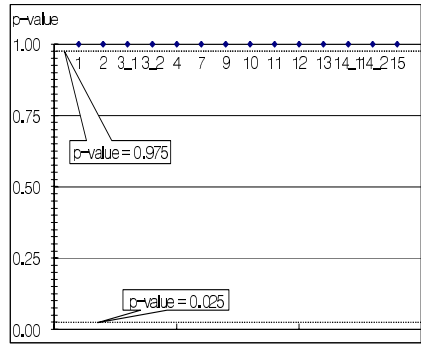
Random properties of the eq. (5) are quite different as shown in Fig. 2. All the results show that they do not pass Diehard, the sequences are not random. The test includes cases for deleting B in eq. (5), by choosing $(a,0)$. From these results we could conclude that the sequence in eq. (4) is random if one of the seeds has full word size. The eq. (5) is not acceptable as a PRNG for any input random seeds.

Table 1. Diehard test items and the nomenclature of data size

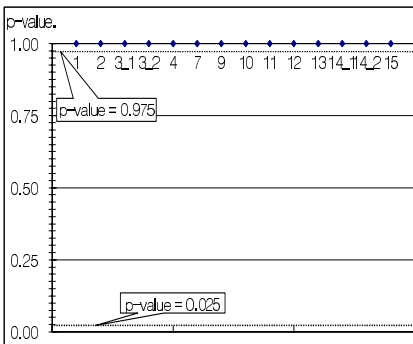
<p>•Diehard Test Item:</p> <p>1: Birthday spacing Test 3-1: Binary Rank test for 31x31 matrices 4: Binary Rank test for 6x8 matrices 9: Parking Lot Test 11: Random Spheres Test 13: Overlapping sums test 14-2: Runs-down Test</p>	<p>2: Overlapping 5-permutation test 3-2: Binary Rank test for 32x32 matrices 7: Count the 1's test on a stream of bytes 10: Minimum distance test 12: The Squeeze Test 14-1: Runs-up Test 15: The Craps Test</p>
<p>•Date Size:</p> <p>(a, b): word size for a and b respectively 1, 2, f: 1-; 2-; full-word-size respectively</p>	



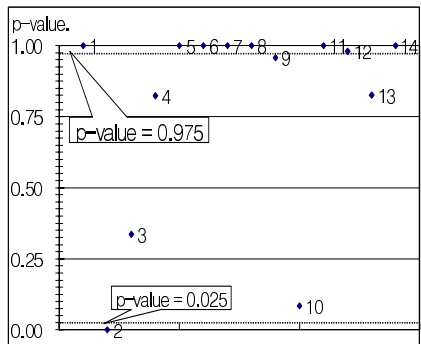
(f, f)



(f, 1)



(1, f)



(1, 1)

Fig. 2. Diehard results for data field sizes of random seed, eq. (5)

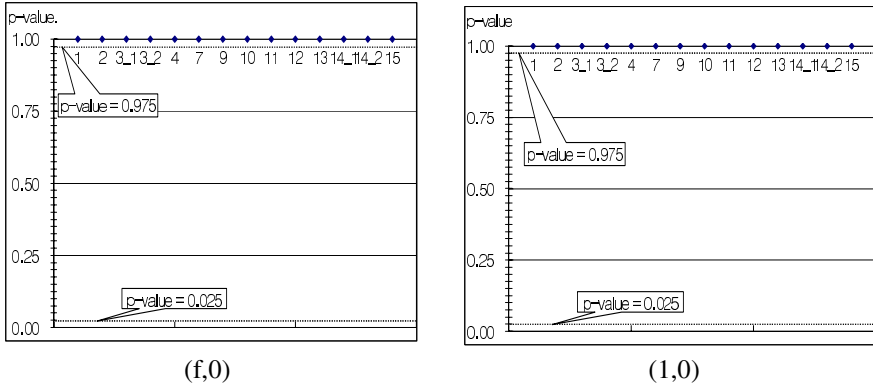


Fig. 2. (continued)

4 Examples

In order to exploit full possibility for eq. (4), we change the bit-size of the random sequence such as {338, 350, 354, 359, 371} by considering optimal normal bases. The random seeds are also changed as before. Another thing is to use other test suite and check various random properties. We add the well-known ENT suite to strength the randomness test since A system that displays high degree of disorder (high entropy)

Table 2. Test results of Diehard suite

Test No.	338-bit	350-bit	354-bit	359-bit	371-bit
1	0.354854	0.651985	0.451643	0.457133	0.856222
2	0.622033	0.032077	0.275088	0.394883	0.187159
3-1	0.970244	0.357225	0.818385	0.495946	0.89977
3-2	0.371633	0.60437	0.430372	0.533298	0.642824
4	0.975782	0.301652	0.323045	0.334351	0.781159
7	0.759872	0.681765	0.706534	0.582362	0.645277
9	0.061532	0.478152	0.10867	0.10728	0.246379
10	0.139288	0.587778	0.538669	0.404724	0.02914
11	0.731812	0.559996	0.865296	0.225171	0.476937
12	0.612573	0.348457	0.439419	0.346584	0.573211
13	0.250332	0.957018	0.403429	0.23069	0.775615
14-1	0.216196	0.318966	0.515968	0.276031	0.798265
14-2	0.418524	0.267296	0.362035	0.811214	0.360601
15	0.468167	0.534111	0.29905	0.5211652	0.569484

Table 3. Test results of ENT suite

Test	Proper Value	338-bits	350-bits	354-bits	359-bits	371-bits
T1	10~90%	50%	75%	75%	20%	25%
T2	8	7.999985	7.999985	7.999986	7.999981	7.999983
T3	127	127.4727	127.4957	127.4817	127.533	127.5
T4	Pi= to within 0.08%	3.144261002 0.08%	3.141295738 0.01%	3.141938177 0.01%	3.141314572 0.01%	3.140908601 0.02%
T5	0	0.000422	-0.000265	-0.00014	-0.000035	0.000357

*T1: Chi Squared Test, T2: Entropy Test, T3: Arithmetic Mean

T4: Pi Estimation, T5: Serial Correlation

Table 4. ENT results for different field sizes

	Chi Squared Test	Entropy Test	Arithmetic Mean	Pi Estimation	Serial Correlation
Seed size	10~90%	8	127	0.08%	0
(1, f)	50%	7.999983	127.5063	0.00%	0.000191
(2, f)	50%	7.999984	127.4803	0.06%	0.000362
(5, f)	90%	7.999986	127.5173	0.06%	0.000215
(f, f)	25%	7.999983	127.4911	0.01%	-0.000135
(f, 1)	75%	7.999985	127.4957	0.01%	-0.000265
(f, 2)	50%	7.999984	127.4909	0.07%	0.000279
(f, 5)	25%	7.999983	127.5178	0.08%	0.000216

can be considered almost or perfectly random, [8]. We create random sequences; in size of the binary file of 11,468,800 megabytes that is expected large enough to check the randomness.

The sample 326-bit random seeds are as following.

$$A = \{0xfd1e0cdc, 0x96124eec, 0xf76df24c, 0xb6554634, 0x31605a0d, \\ 0x19e4ec10, 0x34cf66a2, 0x2284f0c6, 0x94fdd62b, 0x2242337d, 0x2242337d \}$$

$$B = \{0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x2242337d \}$$

Table 2 shows Diehard results for 5 different data sizes. As all the values are in the range of 0.025 and 0.975, we can conclude that randomness does not depend on the size of data. Table 3 is for ENT test results. It gives the same conclusion as before since all the data are in the proper range for all different bits. Table 4 is for different seed sizes. From this table we can see that the eq. (4) gives perfect sequence if one of the seeds is full size.

5 Conclusions

In this paper, we presented a new PRNG with optimal normal basis. Two sequential versions are tested by Diehard suite for different size of random seeds. One of the schemes is shown to be a promising PRNG for various data size, and the other is not random regardless of random seeds. The scheme produced a binary file of 11468800 megabytes and passed both Diehard and ENT suite. The errors for five test items by ENT suite are all negligible. The algorithm is easy to implement and can produce a pseudorandom number sequence of more than 10 32-bit words.

Acknowledgements

This research was supported by the MIC, Korea, under the ITRC support program supervised by the IITA.

References

1. A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray and S. Vo, *A statistical test suite for random and pseudorandom number generators for cryptographic applications*, NIST Special Publication 800-22, <http://www.nist.gov/>, 2001
2. P. Wu, "Random number generation with primitive pentanomials," *ACT Trans. on Modeling and Computer Simulations*, 11, 4, 346-351, 2001
3. J. R. Carr, "Simple random number generation," *Computers & Geosciences*, 29, 1269-1275, 2003
4. L. Lee and K. Wong, "A random number generator based on elliptic curve operations," *Computers and Mathematics with Applications*, 47, 217-226, 2004
5. R.C. Mullin, I.M. Onyszchuk and S.A. Vanstone, "Optimal normal bases in $GF(p^n)$," *Discrete Applied Mathematics*, 22, 146-161, 1988
6. B. Sunar and C.K. Koc, "An efficient optimal normal basis type II multiplier," *IEEE Trans. on Computers*, 50, 1, 83-87, 2001
7. B. Sunar and C.K. Koc, "An efficient optimal normal basis type II multiplier," *IEEE Trans. on Computers*, 50, 1, 83-87, 2001
8. J. Walker, "ENT, A pseudorandom number sequence test program," <http://www.fourmilab.ch/random/>, 1998
9. T. Ritter, "The efficient generation of cryptographic confusion sequences," *Cryptologia*, 12, 5, 81-139, 1991
10. G. Marsaglia and W. W. Tsang, "The 64-bit universal RNG," *Statistics & Probability Letters*, 66, 183-187, 2004
11. G. Marsaglia, "Diehard battery of tests of randomness," <http://www.stat.fsu.edu/pub/diehard/>, 1995
12. B.Ya. Ryabko, V.S. Stognienko and Yu.I. Shokin, "A new test for randomness and its application to some cryptographic problems," *Journal of Statistical Planning and Inference*, 123, 365-376, 2004

Efficient Nonce-Based Authentication Scheme Using Token-Update

Wenbo Shi and Hyeong Seon Yoo

School of Computer Science and Engineering, Inha University,
Incheon, 402-751, Korea
hsyoo@inha.ac.kr

Abstract. In this paper an efficient token-update scheme based on nonce is proposed. This scheme provides an enhancement, resolving some problems with regard to Lee's scheme, which cannot defend against replay and impersonation attacks. Accordingly, an analysis and comparison with Lee's and other schemes, demonstrate that the current paper avoids replay and impersonation attacks, providing mutual authentication, and also results in a lower computation cost than the original scheme.

1 Introduction

User authentication has become an important component of network security, especially over open networks. When a remote user requests a service from a server's, the server is required to authenticate the legitimacy of the user.

To prevent online password eavesdropping, some OTP systems can be considered. However, if the nonce established between user and server is stolen or changed by an attacker, it is possible to fool the server [1, 2]. In 2004, Juang proposed a nonce-based scheme and with the nonce updated at every login instance, however this scheme requires symmetric encryption operations for authentication [3]. This increases the implementation cost for both the user and server. To simplify the authentication process, Chen and Lee also proposed a nonce-based scheme using a hash function, protecting data by using an exclusive-or operation. However the computation cost is still high and it places the burden on the user of generating two good nonce in one login instance [4]. Lee provides a scheme using the Mac address of a LAN card to replace the KIC, in order to strengthen the conventional token update method, however it uses public key cryptosystem for authentication and is not secure against some attacks [5]. Recently, Chen and Yeh proposed a more efficient scheme than Chen and Lee's [4, 6], however, both the user and server are required to generate nonce in one login instance. It is difficult and inefficient for a user to generate an acceptable random number.

In this paper, a more efficient token-update scheme, only requiring the server to generate a random number and establish a new number with a user within a login time, updating the token every time, based on hashing functions and exclusive-or operations, is proposed. In addition, the proposed scheme provides superior efficiency and security.

2 An Efficient Token-Update Authentication Scheme

Lots of schemes use symmetric or asymmetric cryptosystems in authentication [3, 5]. In addition, the use of encryption operations may increase the implementation cost of authentication protocols in both the user and server. Therefore, the proposed scheme uses hash functions and exclusive-or operations to protect the data instead of using cryptosystems. In addition, when a user uses a computer to login, a good nonce can not be generated efficiently. Therefore, in the proposed scheme, the server generates a good nonce and establishes a new nonce every login instance. In this paper, an improved scheme over Lee's scheme is presented. The User (U) and server (S) require three passes to finish.

2.1 Registration Phase

(R.1) $U \rightarrow S: \{M1, M2\}$

Here $M1 = h(token1) \oplus token1$; $M2 = \mathcal{E}_{ks}(token1)$.

U uses his identification (id), password (pw) and Mac address (ma) to construct $token1 = (id \parallel pw \parallel ma)$. U computes $M1$ and $M2$. U sends $\{M1, M2\}$ to S.

(R.2) $S \rightarrow U: \{M3, h(rs)\}$

Here $M3 = h(token1) \oplus rs$.

S gets $token1$ from the $M2$, verifies whether $token1^* = h(token1) \oplus M1$. If they are equivalent, S accepts registration. S generates a random value (rs), computes $M3$ and $h(rs)$. S sends $\{M3, h(rs)\}$ to U, S store $h(token1) \oplus MA$ and $rs \oplus MA$, using his own Mac address (MA) to protect data.

2.2 Login Phase

(L.1) $U \rightarrow S: \{id, M4\}$

Here $M4 = h^2(token1) \oplus rs$.

U provides his id and pw to construct $token1$, extracts $rs = M3 \oplus h(token1)$, then U computes $M4$, sends $\{id, M4\}$ to S.

2.3 Authentication Phase

(A.1) $S \rightarrow U: \{M5, M6\}$

Here $M5 = h(token2) \oplus rs^*$; $M6 = h(token1) \oplus rs^*$.

Upon receiving the message $\{id, M4\}$ in the login phase, S checks $rs^* = h^2(token1) \oplus M4$. If they are equivalent, S generates a new fresh random value rs^* . Next, S construct $token2 = token1 \parallel rs-1$ and computes $h(token2)$, $M5$ and $M6$. S transmits $\{M5, M6\}$ to U.

(A.2) $U \rightarrow S: M7$

Here $M7 = h(token3) \oplus rs-2$.

Receiving message $\{M5, M6\}$, U computes $h(token2) \oplus M3 = h(token1) \oplus M4$. If they are equivalent, U authenticates S successfully. Then, U constructs $token3 = token2 \parallel rs^*$, computes $h(token3)$ and $M7$, sends $M7$ to S. And U stores $h(token1) \oplus rs^*$ and $h(token1) \oplus h(token3)$ for the next login. (Next login time, U constructs $token1$ and computes $h(token1)$ to extract rs^* and $h(token3)$, uses $rs^* \oplus h(token3)$ to login sever).

(A.3) S authenticates U by verifying $M7$

S constructs $token3 = token2 \parallel rs^*$, computes $M7^*$ to verify $M7$. If they are equivalent, S authenticates U successfully.

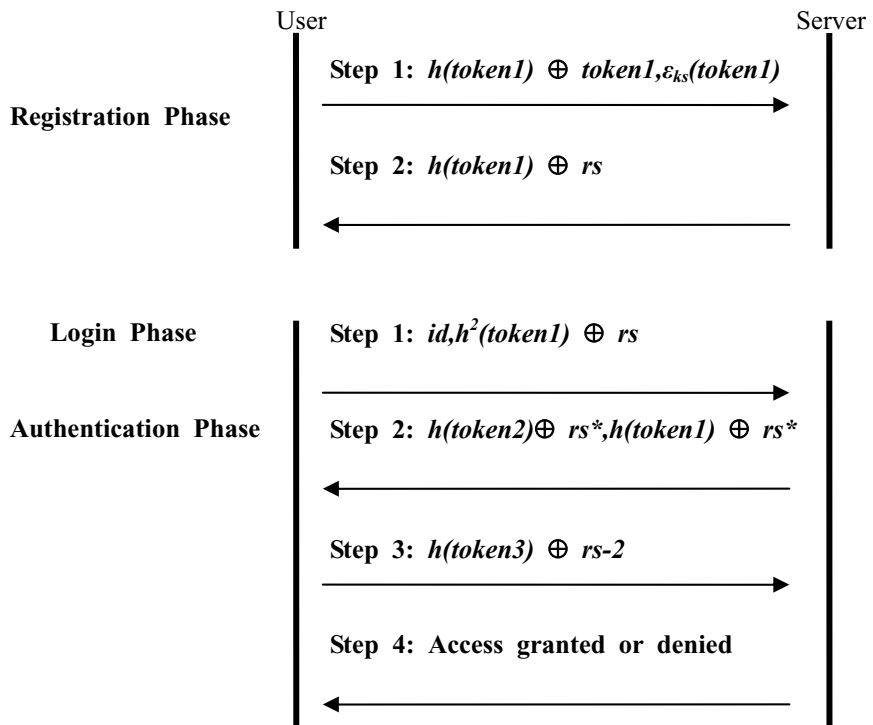


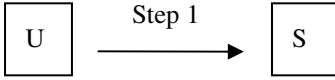
Fig. 1. The proposed scheme

3 Discussion with Lee's [5] and ENA [6] Schemes

A proposed scheme is discussed with Lee's and ENA schemes.

- Proposed scheme : scheme A
- Lee's scheme : scheme B
- ENA schemes : scheme C

3.1 Registration Phase Step 1

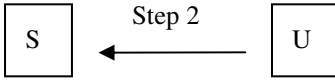


- Scheme A: $\{token1 \oplus token1, \epsilon_{ks}(token1)\}$
- Scheme B: $\{h(token1), \epsilon_{ks}(token1)\}$
- Scheme C: $\{id, pw\}$

In scheme B, U constructs $token1=(id \parallel pw \parallel ma)$, sends $\{h(token1), \epsilon_{ks}(token1)\}$ to S. In scheme C, U sends id and pw to S over a secure communication channel.

Scheme A and scheme B are similar in registration phase step 1, except scheme A use exclusive-or operation to protect data, and scheme B saves token1 directly but scheme A saves $token1 \oplus MA$ to protect $token1$. Scheme C requires a secure communication channel during the registration phase.

3.2 Registration Phase Step 2



- Scheme A: $\{h(token1) \oplus rs, h(rs)\}$
- Scheme B: $\epsilon_{ks}(h(rs))$
- Scheme C: S stores $h(id \oplus x) \oplus pw$ into U' card

In scheme B, S decrypts and computes $h(token1)^*$, if $h(token1)=h(token1)^*$, S accepts registration, returns $\epsilon_{ks}(h(rs))$, and S constructs $token2$, where $token2=token1 \parallel h(rs)$, computes $h(token2)$. In Scheme C, S uses his secret key x and U's id and pw to compute $h(id \oplus x) \oplus pw$, store it on U's smart card.

Scheme A and Scheme C use exclusive-or operation and hash operation, instead of public cryptosystem, in order to protect data, because these operations are more efficient than public cryptosystem from a computation cost aspect. In scheme B, the attacker only requires the user's public key, he can forge data to masquerade as a user, but it is impossible for an attacker to forge the data in scheme A, because the attacker does not have a correct $h(token1)$ to masquerading as a user. The attacker cannot forge data in scheme C either, because the attacker has no correct pw or $h(id \oplus x)$ to forge any valid data.

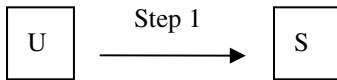
In scheme A and B, user and server establishes $token1$ and a random number rs , is generated by the server. However, in scheme C, a token $h(id \oplus x)$ is established between the user and server.

In scheme B, the user stores encrypted data $\mathcal{E}_{ku}(h(rs))$ protected in the computer by a public key, it requires a user's secret key to decrypt, so it is safe to store it.

In scheme A, the user can store $h(token1) \oplus rs$ in computer, the data requires $h(token1)$ in order to extract rs , and users do not store $h(token1)$, when user wants to login, he constructs $token1$ and computes its hash value in order to obtain rs , therefore it is safe to store it. In scheme C, user stores $h(id \oplus x) \oplus pw$ in the smart card, because an attacker has no correct pw to extracts $h(id \oplus x)$, therefore it is safe to store it.

In scheme B, the server store $h(token2)$ directly. If an attacker steals the $h(token2)$ before the token is updated, the attacker can login using the stolen data. If an attacker changed the $h(token2)$ before the token is updated, the legal user can no longer login. In Scheme A, the server stores $h(token1) \oplus MA$ and $rs \oplus MA$ in the server's computer to protect them.

3.3 Login and Authentication Phase Step 1



Scheme A: $\{id, h^2(token1) \oplus rs\}$

Scheme B: $\{id, h(token2)\}$

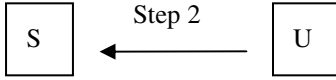
Scheme C: $\{id, pw\}$

In scheme B, U decrypts $\varepsilon_{ku}(h(rs))$ in order to construct $token2$, where $token2=token1||h(rs)$, sends $\{id, h(token2)\}$ to S. In scheme C, U extracts $h(id \oplus x)$ and hash it, generates a random number rc , then computes $h^2(id \oplus x) \oplus rc$ and transmits it to S.

In scheme B, the user concatenates data blocks $token1$ and $h(rs)$ to construct a new token, sends $h(token1||h(rs))$ to the server to login. The data contains $h(rs)$ to ensure it is fresh and comes from a legal user. In scheme A and scheme C, the user protects a random number using a hash function and exclusive-or operation. In scheme A, the user uses $h^2(token1)$ to protect rs . In scheme C, the user uses $h^2(id \oplus x)$ to protect rc . Therefore, this is safe for rs and rc , because an attacker cannot forge a valid rs or rc . However, server authentication can be surpassed in scheme C, by the replay data attack, so this pass does not accomplish server authentication of a user in scheme C. It is difficult and inefficient for a user to generate a good random number, therefore in scheme A and B, a user uses a random number generated by the server and establishes a registration phase.

3.4 Login and Authentication Phase Step 2

In scheme B, S accepts if $h(token2)=h(token2)^*$, update $token2^*$ with rs^* , return $\varepsilon_{ku}(h(rs^*))$ to U. In scheme C, S generates another random number rs and computes $h(h(id \oplus x)||rc) \oplus rs$ and $h(h(id \oplus x)||rc||rs)$ for the user.



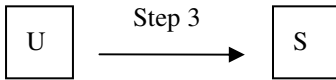
- Scheme A: $\{h(token2) \oplus rs^*, h(token1) \oplus rs^*\}$
- Scheme B: $\varepsilon_{ku}(h(rs^*))$
- Scheme C: $\{h(h(id \oplus x)|| rc) \oplus rs, h(h(id \oplus x)|| rc|| rs))\}$

Schemes A and C use an exclusive-or operation and hash operation in order to protect data instead of a public cryptosystem. In scheme B, an attacker only requires a user’s public key, and can forge fake data to masquerade as a user. In scheme A, because attacker cannot obtain $h(token1)$ or $h(token2)$, a valid rs^* cannot be forged, to masquerade as a user. In scheme C, because the scheme is protected by $h(h(id \oplus x)|| rc)$ and attacker cannot obtain the $h(id \oplus x)$ or rc , an attacker cannot forge a valid rc . In scheme B, the data does not contain any information to inform a user this data is fresh, therefore an attacker can use a replay data attack to masquerade as a user. In scheme A and C, the token contain $rs-1$ or rc to inform the user that the data is fresh, therefore an attacker cannot use a replay data attack to masquerade as a user.

In scheme A, the server uses $h(token2)$ and $h(token1)$ to protect rs^* separately. The valid rs^* is known only by the server, the server must use two tokens to protect rs^* . When user gets two rs^* , if they are equivalent, the rs^* is valid, otherwise it is invalid. If rs^* is correct, a user succeeds in server authentication. Furthermore, if an attacker changes any one of the two data, a user can detect it by comparing two rs^* .

In scheme C, the user authenticates a server by verifying data $h(h(id \oplus x)|| rc|| rs)$ and the server uses this data to protect the random number rs . In scheme B, server saves $h(token2^*)$ directly. In scheme A, the server saves $rs^* \oplus MA$ to protect rs^* .

3.5 Login and Authentication Phase Step 3



- Scheme A: $h(token3) \oplus rs-2$
- Scheme C: $h(h^2(id \oplus x)|| rc+1|| rs+1)$

In scheme C, U transmits $h(h^2(id \oplus x)|| rc+1|| rs+1)$ to S, S calculates $h(h^2(id \oplus x)|| rc+1|| rs+1)$ to authenticate U by verification.

In scheme A, the user uses $rs-2$ to ensure the data is fresh. In scheme C, a user uses $rc+1$ and $rs+2$ to ensure the data is fresh. Scheme A and C add this pass to provide server authentication, and use three-pass to provide mutual authentication. In scheme A, the server stores $h(token3) \oplus MA$ and the user stores $h(token1) \oplus rs^*$ and $h(token1) \oplus h(token3)$. The data is protected by a hash function and exclusive-or operation, therefore it is safe to store it. The security properties of Lee’ scheme and the proposed scheme are summarized in Table 1.

Table 1. Comparison of security properties

	Lee' scheme [5]	Proposed scheme
Replay attack	No	Yes
Impersonating server attack	No	Yes
Password-Guessing attack	Yes	Yes
Man-in-the-middle attack	No	Yes

4 Security and Efficiency Analysis with Other Schemes

Lee's scheme requires public key cryptosystem and does not provide mutual authentication, because the attacker can impersonate the server to cheat a user so [5].

Juang's scheme uses a symmetric cryptosystem and provides mutual authentication in a three-pass data exchange [3, 6]. The user and server establishes a symmetric secret key $k1$ in the registration phase, the user sends nonce $N1$, his id and $\mathcal{E}_{k1}(rc, h(id||N1))$ to the server, the server decrypts data and verifies $N1$, then server transmits $\mathcal{E}_{k1}(rs, N1+1, N2)$ to user, the user decrypts data to authenticate the server. The user creates a new secret key made up of rc , rs , and $k1$, and returning $\mathcal{E}_{k2}(N2+1)$ to the server. The server makes this new key and decrypts data to verify $N2+1$. If this is correct, the server succeeds in authenticating the user.

Chien and Jan's scheme is superior to the previously described schemes, but it requires time synchronization between the user and server. This scheme does not require a verification table or timestamp, and provides mutual authentication [7, 8].

Chen and Lee's scheme does not require a timestamp, and stores a verification table in the server, withstanding a stealing verifier attack, by using three-pass to provide mutual authentication [4]. Firstly, server and user establish a random nonce N , $h^2(pw \oplus N)$ and $h(x||id)$. When a user wants to login, the user transmits his id and another nonce r' to the server. The server generates a new nonce r , and sends user $r \oplus h(x||id)$ and $h(r||r')$. The user authenticates server by $h(r||r')$ and r' . Then the user transmits $c1$, $c2$ and $c3$, to the server, verifies and then authenticates user [4].

ENA scheme provides simple authentication based on a nonce and hash function. It provide mutual authentication within three-pass communications, and do not depend on a cryptosystem [6]. A comparison of the proposed scheme with other schemes is summarized in Table 2.

The efficiency of the proposed scheme is evaluated by comparison with ENA [6] and Lee's [5] schemes. RSA's computation cost can be summarized as a modular exponentiation computation cost, the computation cost of a modular exponentiation is about $O(lnl)$ times. This is compared with a modular multiplication cost in Z_n^* , random-number generation and hashing can be ignored [8]. In addition, the exclusive-or operation can be performed very efficiently and the extra computation cost is negligible.

The computation cost of Lee's scheme for login and authentication is asymmetric key encryption, asymmetric key decryption, three hashing operations and a random

number generation. Therefore, the ENA scheme and the proposed scheme are more efficient than Lee’s scheme.

The computation cost of ENA scheme for login and authentication is nine hashing operations, two random numbers, and five exclusive-or operations. And the computation cost of the proposed scheme for login and authentication represents five hashing operations, one random number, and fourteen exclusive-or operations. The computation cost of the both schemes is extremely low. The comparisons of computation costs are summarized in Table 3.

Table 2. Comparisons with some others schemes

	Encryption	Verification table	Timestamp	Mutual authentication
Lee Joungho [5]	Yes [5]	No [5]	No [5]	No
Juang [3]	Yes [3]	No [3,6]	No [3,8]	Yes [3,6,8]
Chien and Jan [7]	No [6,7,8]	No [6,7]	Yes [6,7,8]	Yes [6,7,8]
Chen and Lee [4]	No [4]	Yes [4]	No [4]	Yes [4]
Chen and Yeh [6]	No [6]	No [6]	No [6]	Yes [6]
Our scheme	Yes	Yes	No	Yes

Table 3. Comparisons of computation costs of number functions

	Lee’s [5] scheme		ENA [6] scheme		proposed scheme	
	user	server	user	server	user	server
Registration phase	1T(e)	1T(d)			1T(e)	1T(d)
	1T(h)	1T(e)	0T(h)	1T(h)	1T(h)	1T(h)
		3T(h)				1T(r)
		1T(r)		1T(⊕)	1T(⊕)	5T(⊕)
Login and authentication phase	1T(d)	1T(e)	4T(h)	5T(h)	3T(h)	2T(h)
	1T(h)	2T(h)	1T(r)	1T(r)		1T(r)
		1T(r)	3T(⊕)	2T(⊕)	8T(⊕)	6T(⊕)

T():computation time; h: secure one-way hash; r: generate a random number; e: encryption; d: decryption; ⊕:exclusive-or

5 Conclusion

In this paper, a new efficient scheme using a hash function and exclusive-or operation is presented, to isolate replay attack and impersonation attacks. In addition, the scheme provides mutual authentication, and only requires a nonce value generated by server and establish a new number for a user during login. According to the analyses in section 3, the proposed scheme overcomes the security drawbacks of Lee’s and provides improved security. Compared with other schemes described in section 4, the proposed scheme shows good security and efficiency.

Acknowledgements

This research was supported by the MIC, Korea, under the ITRC support program supervised by the IITA.

References

1. N.M. Haller, "The S/Key (TM) one-time password system," ISOC Symposium on Network and Distributed System Security, 151-158, 1994.
2. B. Soh and A. Joy, "A Novel Web Security Evaluation Model for a One-Time- Password System," IEEE/WIC, IEEE Computer Society, WI('03), 413-416, 2003
3. W.-S. Juang, "Efficient password authenticated key agreement using smart card," Computers & Security 23 (2004), pp. 167-1738
4. T.H. Chen, W.B. Lee and G. Horng, "Secure SAS-like password authentication schemes," Computer Standards & Interfaces, 27, 25-31, 2004
5. Joungho Lee, Injoo Jang and Hyeong Seon Yoo, "Modified token-update scheme for site authentication," LNCS 3481,111-116, 2005
6. Yen-Cheng Chen,Lo-Yao Yeh, "An efficient nonce-based authentication scheme with key agreement," Applied Mathematics and Computation, 169, 982-994, 2005
7. H.Y. Chien, J.K. Jan and Y.M. Tseng, "An efficient and practical solution to remote authentication with smart card," Computers & Security, 21, 372-375, 2002
8. Chun-I Fan, Yung-Cheng Chan and Zhi-Kai Zhang, "Robust remote authentication scheme with smart cards," Computers & Security , 2005

An Efficient Management of Network Traffic Performance Using Framework-Based Performance Management Tool

Seong-Man Choi, Cheol-Jung Yoo, and Ok-Bae Chang

Dept. of Computer Science & Statistical Information, Chonbuk National University,
664-14, 1Ga, Duckjin-Dong, Jeonju, Jeonbuk, 561-756, South Korea
{sm3099, cjyoo, okjang}@chonbuk.ac.kr

Abstract. As the network-related technology develops the number of both internet users and the usage are explosively increasing. The networking traffic is increasing in the campus as the networking system inside universities, following the trend, adds more nodes and various networking services. Nonetheless, the quality of services for users has been degraded. Accordingly, core problems, which can cause troubles for network management, design and expansion of the network, and the cost policy, have developed. To effectively cope with the problems an analysis of a great number of technicians, tools, and budget are needed. However, it is not possible for mid and small-sized colleges to spend such a high expenditure for professional consulting. To reduce the cost and investment of creating an optimized environment, analysis of the replacement of the tools, changing the network structure, and performance analysis about capacity planning of networking is necessary. For this reason, in this paper, framework-based performance management tools are used for all steps that are related to the subject of the analysis for the network management. As the major research method, the current data in detailed categories are collected, processed, and analyzed to provide a meaningful solution to the problems. As a result we will be able to manage the network, server, and application more systematically and react efficiently to errors and degrading of performance that affect the networking tasks. Also, with the scientific and organized analyses the overall efficiency is upgraded by optimizing the cost for managing the operation of entire system.

1 Introduction

Sudden increase of nodes and various multimedia contents of university network increase traffic and decreases service quality. Universities which have a small and medium sized network have problems with operation, error management and performance of the network and it is difficult to continue professional consulting on the network due to a lack of enough people who can control it, high price and difference of a view point [1, 2]. Performance analysis is necessary to change the configuration, to alter the equipment and to plan the capacity more efficiently to prevent excessive investment in university network and establish optimized network environment. The purpose of this paper is to find an analysis method using framework-based Integrated Management System (IMS) and to offer solutions for the problems induced by the analysis results.

The background of a theory of performance analysis is described and merits of IMS and configuration for integrated management are examined in terms of their functions. This method has been applied to network traffic performance analysis of JEUS(Jeonju national university of Education Ultramodern System) and performance analysis results and problems are induced. The method aimed to select important targets which affect the performance and collect, process and analyze data of detailed items using IMS. A method which analyzes relation between each item and their relation with targets in detail has been tried. Main analysis system is InfRanger™ of IMS and assistant tools are NetworkHealth™, MRTG™(Multi Router Traffic Grapher) and Sniffer™ which are network management tools to collect and analyze various information. Management protocol of each unit or agent software are used to collect system inter information such as server and router [3, 4].

This paper has been organized as follows: Section 2 discusses related works: introduction of framework - based IMS, its merits and configuration module. Section 3 discusses framework-based performance management tool. Section 4 provides a summary of network traffic performance analysis in the enterprise environment: targets of network traffic performance analysis in the enterprise environment, performance analysis procedure, configuration of selected targets and selection of the targets for performance analysis. Section 5 outlines network traffic performance simulation analysis results in the enterprise environment : private lines, routers, Web servers, application response time. Section 6 is about conclusions and future works.

2 Related Works

Currently network traffic performance management analyzes quantity, sorts, and specific characters of the traffic which comes into and out from routers and gateways when an object is the unit network or the internet, gets the information on the network performance and acquires its error possibility [4, 5]. When the efficiency and coefficient of utilization are informed to a manager then the manager manages the network performance based on the information [6]. Section 2 is on the tools for network traffic performance management.

2.1 SNMP

SNMP(Simple Network Management Protocol) is a simple network management protocol which consists of TCP/IP and is used mostly now. It can be explained with a SNMP agent/management station model. SNMP agent is set up to a system of a management object and provides MIB(Management Information Base) information of the management equipment to the management station [4, 7].

SNMP uses UDP(User Datagram Protocol), a transport layer which is the 4th layer of OSI 7 model, so there is no need to maintain the connection between the agents and the management station to transmit and receive messages. Therefore it doesn't use a lot of resources, which can be a merit of SNMP, but the message exchange between the agents and the management station is not reliable [8]. SNMP is not efficient enough to handle an extensive amount of table information. It is possible to minimize request time and collection time of management when the network management station takes packets of a sub-network in real-time and monitors information flows [9].

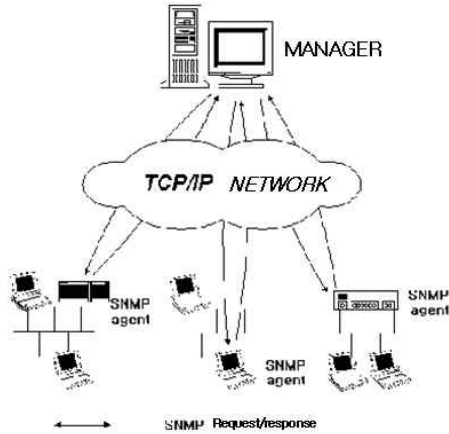


Fig. 1. SNMP based network management system

2.2 TMN

TMN(Telecommunications Management Network) is a systematic structure which supports various forms of operation systems with OSF(Operations System Functions) and communication equipment to exchange management information by using standardized protocols and interfaces [7, 10]. It intends to be connected with each other, reused and standardized of network resources based on object-oriented technology. Figure 2 shows the structure of TMN system.

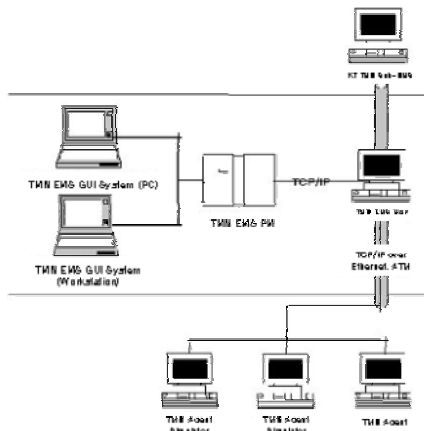


Fig. 2. TMN system

TMN manages network with a method of standardized information exchange between a manager of the management system and agents in a management object. The manager uses the standardized protocol and gives an order to the agents to acquire information on the management object or transmit a management order. The agent takes the order, gives indication for a proper action to the management object and reports the result to the manager [11]. TMN system should connect proper number of agents for one manager and operate them based on the information quantity of the messages exchanged between the manager and the agents.

2.3 Consideration of the Related Works

Many performance analysis tools of network traffic automatize information collection and analysis on operation state. They watch the operation state of the network constantly and give the state of coefficient of utilization and fusibility for the network equipment and the circuit and performance information. The performance analysis tools of network traffic are developed according to the equipment on the network or an information collection method on the traffic.

There are 2 types of the analysis tools [4]. First, when they are used they are connected to specific segments of the network and have a lookout ability for detecting problems. To watch all the segments analysis equipment must be established for each segment and it requires a high overhead cost. So it is not a suitable method for network analysis, but when the network has problems they are the best tools to solve them by concentrated analysis. Second, they are respectively simple and expense is low. They collect data from agents on the network equipment such as hub, bridge, and router and analyze them. The method can be used to analyze the network constantly and to construct an economical analysis system so it is the most widely used analysis method on the general network. The merits of SNMP are that it doesn't require a constant connection to transmit and receive messages between agents and the management station and use a lot of resources. However its message exchange between agents and the manager is not reliable and it inefficiently absorbs table information so it is not a suitable performance analysis tool for JEUS network traffic. And TMN has a very strong query ability of management information with scoping and filtering but it is slow and difficult to use.

Therefore the InfRangerTM, framework-based integrated management system used in this paper is a Web-based structure, accesses anywhere and anytime and make various statistic analysis and various forms of reports as a manager wants. And it provides objective data for the system extension and design with its ability of statistic collection and present state analysis for every performance to make general analysis and integrated management possible for the network, the server and the application. Therefore it is used as a performance analysis tool for network traffic of JEUS with professional consulting service.

3 Synopsis of Framework-Based Performance Management Tool

The InfRangerTM is framework-based IMS because it consists of NMS, SMS and PMS(PC Management System), CMS and SLP(Application Service Level Performance) or AMS. The InfRangerTM collects information from management

target equipment using SNMP as a network management protocol to manage and analyze network. SNMP consists of SNMP agent and SNMP manager. The manager gets MIB on management equipment from the agent. SNMP uses commands such as get, set and get - next and gives management information service to application program [4]. The InfRanger™ collects information mainly using SNMP from MIB-II. Each group of MIB-II has a lot of items related to performance. Table 1 shows the contents of interface group related to network performance analysis [4, 8].

Table 1. Construction of interface group

Object Identifier(OBI)	Contents
ifInOctets	Number of bytes received by interface
ifOutOctets	Number of bytes transmitted from interface
sysUpTime	Time after re-initialize network management targets for the last time
ifType	Interface types distinguished by protocols of physical/data link class
ifInUcastPkts	Number of unicast packets transferred to upper class protocol
ifInNUcastPkts	Number of non-unicast packets transferred to upper class protocol
inOutUcastPkts	Number of transferred packers including deserted upper class protocols and non-transferred packets to unicast address
ifInDiscards	Number of deserted packets which are not transferred to upper class with buffer overflow
ifOutDiscards	Number of packets non-transferred to server network address
ifInError	Number of packets which have errors input through interface
ifOutError	Number of packets which have errors output through interface

4 Summary of Network Traffic Performance Analysis in the Enterprise Environment

Traffic performance analysis procedure and contents of the selected targets are shown in this section. Targets of performance analysis of JEUS in the enterprise environment are selected.

4.1 Performance Analysis Procedure and Structure of the Selected Targets

Network traffic performance analysis procedure in the enterprise environment consists of 4 phases Figure 3. The first two phases are supporting and must be

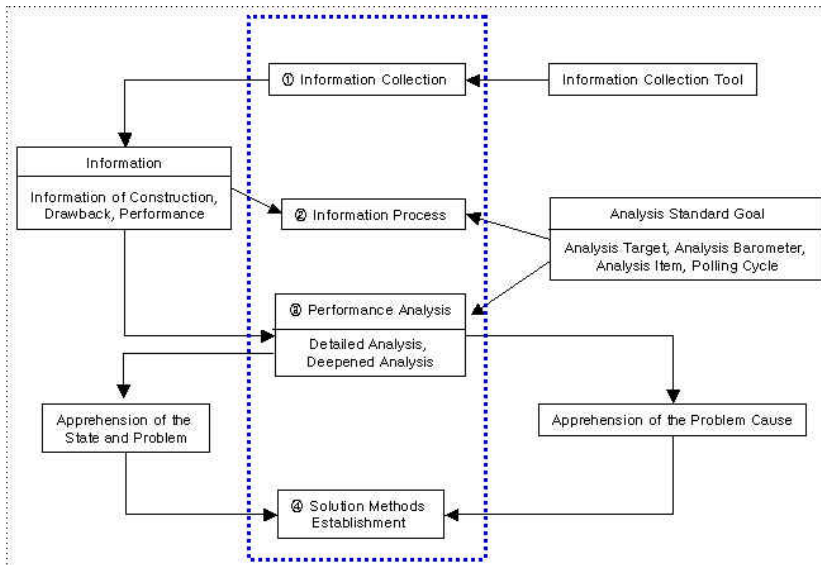


Fig. 3. Procedure of network traffic performance analysis in an enterprise environment

proceeded steps and the last two phases are the steps which make the utility of performance analysis high by giving solutions for problems.

Various types of analysis information are received through user interface of automatic collect tool and each device at information collect phase Figure 3. Collected information is processed in various types reflecting analysis targets, analysis guide posts, analysis items, polling cycles and service goals at information process phase. Performance analysis phase consists of detailed analysis and deepening analysis. Each analysis item is analyzed in detail by applying analysis criterion in various view points at detailed analysis phase. State of the problem and degree of the condition but a cause can be found at detailed analysis phase. Interrelation between analysis attribute, interrelation between different analysis targets and items, state of the problem and its cause are found at deepening analysis phase. Analysis information is analyzed deeply in the interrelation and identifies the cause of the problem. Solution for the cause of the problem which is found at detailed analysis and deepening analysis is reflected to make a decision at solution establishment phase.

The structure of JEUS network applied to network traffic performance analysis in the enterprise environment is shown Figure 4. The backbone of campus network consists of 1,000Mbps and high-ethernet at some intervals for distance limit. Backbone switch equipment interval is triangle shape and supports a circuit way for errors. The backbone between gigabit switches covers 550m with 1000BASE-SX in 50 μ /125 multi-mode fiber. UTP cable includes category 5, category 5E and category 6. Its switching hub at the edge gives network service to clients. It connects with the main server directly and guarantees bandwidth of 100Mbps. It uses 6 authorized IPs of C class to identify 700 nodes of campus network.

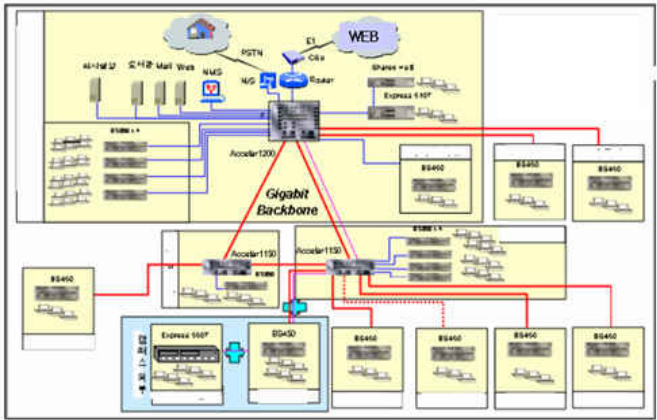


Fig. 4. Network construction of JEUS

4.2 Selection of Performance Analysis Targets of JEUS in the Enterprise Environment

A solution for mid and long-term development plan will be given a measuring network health chart and analyzing traffic tendency of backbone intervals of JEUS in the enterprise environment. Status of contents of JEUS is grasped and coefficient of utilization, coefficient of process, traffic amount, useable protocols and response time are analyzed. Selection of performance analysis targets of JEUS in the enterprise environment is presented in Table 2.

Table 2. Target selection of network traffic performance analysis in an enterprise environment

Analysis Targets	Analysis Items
Private Line (Konet E1)	Coefficient of utilization, coefficient of packet loss
Router (Cisco 3640)	Coefficient of memory/cpu use, pps, coefficient of packet transmission loss, coefficient of error, coefficient of operation
Web Server (Sun 450)	Coefficient of cpu/memory use, coefficient of swap use, packet analysis, status of process, coefficient of network use, Web performance
Application Response Time	Backbone interval and user terminal device interval/protocol response time

5 Network Traffic Performance Simulation Analysis Results in the Enterprise Environment

Analysis systems which are applied in this section are NMS, SMS and SLP of the InfRanger™ as main systems and network health, MRTG™, Sniffer™ and related

commands as assistant tools [2, 10]. The InfRangerTM gives critical value which is applied to set an analysis criterion. The order of analysis is private line performance simulation analysis result, router performance simulation analysis result, Web server performance simulation analysis result and application response time performance simulation analysis result.

5.1 Private Line Performance Simulation Analysis Result

Private line is inter-networking which gives an integrated LAN environment of JEUS. The items of performance analysis targets are coefficient of line utilization and coefficient of packet loss. Performance analysis tool, MRTGTM, measures and analyzes serial ports of routers [2, 11]. Traffic is input and output daily, weekly, monthly and yearly. Max value, Average value and Min value are given in the form of Web. Traffic is measured in bps (bit per second) unit and Bps(Byte per second) unit by MRTGTM version.

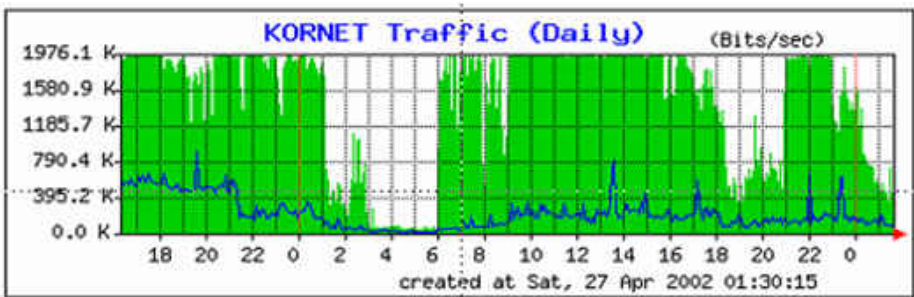


Fig. 5. Performance simulation analysis results of private line

The coefficient of utilization Figure 5 is calculated as follows: Bandwidth is E1 (2.048), Traffic 877,323,886 Byte for 1 hour, the coefficient of utilization is $(877,323,886/921,600,000) \times 100 = 95.00\%$. Daily graph of traffic measurement shows max value, average value and current value of input and output traffic in bps Figure 5. The horizontal of the graph is time and the vertical is traffic. The criterion of the vertical traffic is measured max traffic and the value is line capacity (1976.4kbps).

Private line performance simulation analysis result, which comprises 94% of the bandwidth of the line used during business hours, presents a serious bottleneck situation. The backbone of JEUS is gigabit Ethernet and high-Ethernet to terminal so the total performance of LAN intervals is generally good but the input traffic from outside is large. This is the source of many complaints among university users. Output traffic average, 10%, shows insufficient information service of the university. Construction of a service system which can offer various contents is needed as a solution.

5.2 Router Performance Simulation Analysis Result

Router is an inter-networking equipment which connects between LAN without relation to network construction or protocols. The router can get more traffic than any

other communication equipment; therefore, Cisco 3640 model has been selected as a performance analysis target and performance related items have been analyzed to reflect future network operation [12]. The model is module type multi-function access platform which supports sound/data integration, VPN(Virtual Private Network), dial access, multi-protocol data routing, 4 network module slots and 50~70Kpps performance. The items of performance analysis are coefficient of memory utilization, coefficient of CPU utilization, coefficient of transmission failure, pps(packet per second), coefficient of error, coefficient of operation, waiting time and coefficient of collision of router. They are analyzed using router commands and IMS. Data collected for 30days are analyzed and the simulation result is shown at Table 3. Coefficient of memory utilization expresses average value and max value of memory use of current using router equipment. Total memory is calculated as free memory 100/total memory using information of MIB-II. Coefficient of transmission failure is failure of packet transmission per certain time at router and PPS is the number of packets per a second which are transmitted to router equipment.

Table 3. Performance simulation analysis results of routers

Statistic of Coefficient of Router Utilization								
Class		Memory Utilization(%)	CPU Utilization(%)			Packet Transmission Failure(%)	PPS	
Average		36.4	4.03			0.07	462.00	
Max		38.78	8.50			1.37	580.00	

Status of Router Network Performance								
Response Time(ms)			Coefficient of Error(%)			Average Data(bps)		
Current	BL	Comparison	Current	BL	Comparison	Current	BL	Comparison
0.03	0.15	-0.11	0	0	0	2.91M	3.59M	-681.76K

The coefficient of memory utilization measured by max value criterion is under 40% of the critical value, which is acceptable, and the coefficient of CPU utilization is below 40% of critical value, which is excellent. The coefficient of packet transmission failure is under 0.1% which is acceptable, and PPS is 462pps much less than 50-70Kpps of Cisco router performance which is excellent. Therefore router performance analysis simulation result is generally good and elements of performance degradation were not found.

5.3 Web Server Performance Simulation Analysis Result

A Web server is one of the servers which have a lot of network traffic in the enterprise environment. It is used as a basic data to analyze physical analysis and coefficient of

utilization and establish capacity plan. System content collected by InfRangerTM related to current performance of Sun450 Web server is presented in Table 4.

Table 4. Construction state of Web server systems

Class	Contents Information	Class	Contents Information
Operating System	Sun 5.6	Network Interface	100Mbps
Number of CPU	1	Whole Capacity of the Disk	16,495MB
Whole Memory	512MB	Whole Swap	1,538MB

The items of Web server performance analysis outlined in Table 5 are coefficient of CPU and memory utilization, coefficient of swap area utilization, current use situation of network, process, Web server performance and coefficient of file system use. The Web server collects and analyzes information which is related to the items.

Table 5. Performance simulation analysis results of Web server

Class	CPU(%)	MEMORY(%)	SWAP(%)
Average	32.549	97.33	41.02
Max	54.15	98.68	46.47
Excess	0	0	0

The coefficient of CPU utilization is as high as 54% on general days and the coefficient of memory utilization is over critical value of average coefficient of memory utilization. The coefficient of disk utilization, coefficient of network utilization, packet error and collision packet of capacity plan are stable but extra space of some file systems is not enough. There can be errors which a server can not catch and cause performance degradation so a user should operate a Web server carefully and construction of continuous observation is needed.

5.4 Application Response Time Performance Simulation Analysis Result

One of the most frequent complaints of users is delay of response in using the network. Over 90% traffic of JEUS is from out-of-campus so response time analysis for this is essential. Measurement on backbone interval response time, terminal interval response time and application response time is needed. Information of application response time can be collected by measuring the protocol of a service port. Response time is analyzed

as an item of performance analysis and framework-based IMS uses NMS and SLP of InfRangerTM. SnifferTM, protocol analysis tool is used as an assistant management system to compare the result. Application response time performance simulation analysis result of every server and backbone interval is provided in Table 6.

Table 6. Performance simulation analysis results of application response time by each server/backbone section

Class	Server 1	Server 2	Server 3	Router Serial	Backbone
Average	0.6	0.7	0.6	0.4	1.2
Max	16	16	16	16	10
Min	0.1	0.1	0.1	0.1	0.1
Total number of polling	756	755	442	2487	931

The total number polling provided in Table 6 is the number of performance of ping. The measurement values of server group at an integrated management server are maximum value, 16ms and average maximum value, 1.2ms which are acceptable. Backbone intervals have an average maximum value, 10.5ms and critical value, 10ms which are acceptable, but maximum value, 188ms which is over critical maximum value, 100ms shows instability of packet flow between the intervals with quality of the line, communication equipment and coefficient of utilization.

Application response time performance simulation analysis result reveals that response time of servers and backbone intervals exceed critical value because the length of the packet flow between specific intervals was instable but stable between general intervals. Protocol measurement of out-of-campus Web server of JEUS exceeds critical time at response time and solubility and in-the-campus server shows fine state. Elements of network speed degradation by computer virus should be known and construction of computer virus protection system is required to prevent delays in future application response time. It is necessary to reconstruct the network as defined as port-based VLAN(Virtual LAN) to take hold of unnecessary traffic flow.

6 Conclusions and Future Works

Network traffic performance analysis methods in the enterprise environment are presented comprehensively and network information of JEUS is collected more easily using framework-based IMS then carry out performance analysis scientifically and in structure. Framework-base IMS is a system which measures, analyzes and expects every item related to management function, analysis function and management and

analysis targets for network management and analysis.

The merits of using framework-based IMS are : First, various resources such as network, server, application, storage and database can be managed in structure at framework-based platform. Second, it provides the convenience of management with Web-based interface. Third, it supports the performance of management in functions such as resource management, error management, configuration management, performance management, security management and report management. Fourth, it makes information collection of detailed status and performance of network resources easy and performance analysis targets and performance analysis items express in various. Fifth, it makes price optimization of operation management of total information systems and enhances management quality by normalizing network management structure. Technical analysis methods of performance analysis critical value which can not be studied because of various performance analysis targets or methods and solutions for the problems derived form analysis results must be investigated as part of future research.

References

1. Seong-Man Choi, Wan-Seob Byoun, Cheol-Jung Yoo, Yong-Sung Kim, Ok-Bae Chang and Gyu-Yeol Tae, Network Traffic Performance Analysis using Framework-based Integrated Management System, in Proceedings of The 30th KISS Spring Conference(B), Vol. 30, No. 1(2003) 145-147
2. Seong-Man Choi, Gyu-Yeol Tae, Cheol-Jung Yoo and Ok-Bae Chang, An Efficient Management of Network Traffic using Framework-based Performance Management Tool, in Journal of KISS:Computing Practices, Vol. 11(2005) 224-234
3. Seong-Ho Cho and Chung-Suk Kim, A Review of Measurement/Analysis and Management Method of Network Traffic, POWER ENGINEERING, Vol. 8, No. 3(1997)
4. InfRanger(NMS, SMS, SLP) : <http://www.kdcorp.co.kr>, <http://infranger.com>
5. Tat Chee Wan, Alwyn Goh, Chin Kiong Ng and Geong Sen Poh, Integrating Public Key Cryptography into the Simple Network Management Protocol (SNMP) Framework, in Proceedings of TENCON, Vol. 3(2000) 271-276
6. Nishiyama S, Ono C, Obana S and Suzuki K, Distribution Transparent MIB based on MSA(Management System Agent) Model, in Proceedings of Parallel and Distributed Systems(1996) 478-485
7. NMS : <http://www.inti.co.kr/products/products.html>
8. Liebeherr J and Tai A, A Protocol For Relative Quality-of-Service in TCP/IP-Based Inter Networks, in Proceedings of Architecture and Implementation of High Performance Communication Subsystems(1995) 62-65
9. Ansari F and Acharya A, A Framework for Handling Route Changes and Aggregation in IPSOFACTO, in Proceedings of Global Telecommunications Conference(1998) 3751-3756
10. Cisco Systems, Inc., Software Configuration Guide For Cisco 3600 Series Routers
11. CISCO : <http://www.cisco.com/kr/>
12. MRTG, SLP : <http://www.gcc.go.kr/biz/biz.asp>

A Prediction Method of Network Traffic Using Time Series Models

Sangjoon Jung¹, Chonggun Kim², and Younky Chung¹

¹ School of Computer Engineering, Kyungil University,
712-701, 33 Buho-ri, Hayang-up, Gyeongsan-si, Gyeongsang buk-do, Korea
sjjung@kiu.ac.kr,
ykchung@kiu.ac.kr
<http://www.kiu.ac.kr>

² Dept. of Computer Engineering, Yeungnam University,
712-749, 214-1, Dae-dong, Gyeongsan-si, Gyeongsangbuk-do, Korea
cgkim@yu.ac.kr
<http://nety.yu.ac.kr>

Abstract. This paper describes a method to derive an appropriate prediction model for network traffic and verify its trustfulness. The proposal is not only an analysis of network packets but also finding a prediction method for the number of packets. We use time series prediction models and evaluate whether the model can predict network traffic exactly or not. In order to predict network packets in a certain time, the AR, MA, ARMA, and ARIMA model are applied. Our purpose is to find the most suitable model which can express the nature of future traffic among these models. We evaluate whether the models satisfy the stationary assumption for network traffic. The stationary assumption is obtained by using ACF(Auto Correlation Function) and PACF(Partial Auto Correlation Function) using a suitable significance. As the result, when network traffic is classified on a daily basis, the AR model is a good method to predict network packets exactly. The proposed prediction method can be used on a routing protocol as a decision factor for managing traffic data dynamically in a network.

1 Introduction

With a rapid growth in Internet technology, network traffic is increasing swiftly. An increase in traffic has a large influence on the performance of a total network. Therefore, management of traffic is an important issue of network management[1][2][3]. There are many methods of evaluating network performance for users to rely management results. In order to achieve proper results, analyzing traffic is also important[4]. Analyzing network traffic is to alert to the system manager who prepares an invasion of network and reacts a proper behavior when the system is willing to collapse[5][6].

Monitoring traffic consists of gathering and investigating transferred packets. It provides additional information to the manager through analyzed results. This information is provided with extending forms such as the amount of packets, traffic sources, congestion positions, types of traffic, and the maximum amount of traffic and so on. An analysis of Traffic gives information to the manager while Internet services

are provided continuously[7][8][9]. There are two monitoring systems that use management information. One method is a management system to use the SNMP, and the other is a monitoring system to monitor real-time packets[4][10][11]. In order to manage network performance efficiently, it is important to not only analyze types of traffic but also predict the amount of traffic. Predicting network traffic allows us to act properly before congestion occurs[5][6][9].

In this paper, we demonstrate that time series models can be used in predicting network traffic. In order to do this, the time series model must be checked before prediction. The satisfaction of the time series model is checked by the stationary assumption whether it has stationarity or not. The stationary assumption can be evaluated by using ACF and PACF.

2 Prediction Models and Related Works

2.1 Time Series Models

A time series is a sequence of observations that are ordered in time (or space). If observations are made on some phenomenon throughout time, it is most sensible to display the data in the order in which they arose, particularly since successive observations will probably be inter-dependent[12][13]. Time series data analyzed according to the function can be made into a model by way of an analysis method. Time series models are as follows[6][12][13].

2.1.1 The AR(Auto Regressive) Model

The auto regressive approach is based on the premise that each observation in a time series is related in a consistent and identifiable way to one or more previous observations of the same series. The form of this model is as follows.

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \dots + \phi_p Z_{t-p} + a_t \quad (1)$$

Here, the white noise is $a_t = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z_t^2}{2\sigma^2}}$.

2.1.2 The MA(Moving Average) Model

The moving average is a form of average that has been adjusted to allow for seasonal or cyclical components in a time series. The function of this model is to smooth the original time series by averaging a rolling subset of elements from the original series, consisting of an arbitrary selection of consecutive observations. The moving average process can be thought of as the output from a linear filter with a transfer $\theta(B)$, when a white noise is inputted.

$$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q} \quad (2)$$

2.1.3 The ARMA(Auto Regressive Moving Average) Model

To obtain an accurate model, the inclusion of both AR and MA terms is sometimes necessary. The form of this model is as follows.

$$\begin{aligned}
 Z_t' &= \phi_1 Z_{t-1}' + \phi_2 Z_{t-2}' + \dots + \phi_p Z_{t-p}' + a_t \\
 &\quad - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q} \\
 \phi_1, \theta_1, a_t &: \text{A value of estimation}
 \end{aligned}
 \tag{3}$$

An autoregressive model of order p is conventionally classified as $AR(p)$, while a moving average model with q terms is classified as $MA(q)$. Thus, a combination model containing p autoregressive terms and q moving average terms is classified as $ARMA(p,q)$.

2.1.4 The ARIMA(Auto Regressive Integrated Moving Average) Model

The process of ARIMA modeling allows the two modeling approaches to be integrated. If the object series is differenced d times to achieve stationarity, the model is classified as $ARIMA(p,d,q)$, where the symbol "I" signifies "integrated." In theory, ARIMA models are the most general class of models for forecasting a time series that can be stationarized by such transformations as differencing and logging[12]. The form of this model is as follows.

$$\begin{aligned}
 Z_t &= \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \dots + \phi_p Z_{t-p} + u_t \\
 &\quad - \theta_1 u_{t-1} - \theta_2 u_{t-2} - \dots - \theta_q u_{t-q} \\
 \phi(L)z_t &= (L)u_t \quad \text{even if,} \quad \phi(L) = 1 - \phi_1 L - \dots - \phi_p L^p \\
 &\quad \theta(L) = 1 - \theta_1 L - \dots - \theta_q L^q \\
 \text{Therefore, } \phi(L)\nabla^d y_t &= (L)u_t
 \end{aligned}
 \tag{4}$$

2.2 ACF(Autocorrelation Function) and PACF(Partial Autocorrelation Function)

An inference process is necessary to execute prediction with time series models[12]. The stationary assumption is regarded as a previous step in the prediction process in order to make an accurate prediction. Stationary time series data have a constant mean, a constant variance, and the covariance is independent of time. An assumption of stationary data is essential for predicting network traffic[12][13]. Non-stationary data differ from stationary data in the process of probabilities. Stationary data have a constant mean and a constant variance at particular time. Otherwise, non-stationary data has not. The stationary assumption can be evaluated by using ACF and PACF[12][13]. We can obtain the results when these two functions satisfy the stationary assumption. An autocorrelation refers to the correlation of time series with its own past and future values. The autocorrelation function is a tool in assessing the degree of dependence in recognizing what kind of the time series model follows. When we try to fit a model to an observed time series data, we use the function based on the data. We can obtain the equation of the function as follows.

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k} (Z_t - \bar{Z})(Z_{t+k} - \bar{Z})}{\sum_{t=1}^n (Z_t - \bar{Z})^2}
 \tag{5}$$

The partial autocorrelation means that correlation between observations Z_t and Z_{t-k} after removing the linear relationship of all observations between from Z_{t-1} to Z_{t-k+1} . The PACF shows the added contribution of Z_t to predict Z_{t+1} . The equation can be obtained as follows.

$$\phi_{kk} = \text{Corr} [Z_t, Z_{t-k} | Z_{t-1}, Z_{t-2}, \dots, Z_{t-k+1}] \quad (6)$$

In order to verify the AR model, we obtain the results of two functions that have significances. In aspect of ACF, as the order is increasing, the result of the function decreases exponentially. Otherwise, PACF has a different characteristic. The result of the function has a significant value when the lag is k , while the function has no significance at the rest of lags[12][13].

The characteristic of the MA model, both ACF and PACF result have different results in contrast to the AR model. The PACF has a significant value when the lag is k . Otherwise, the ACF has a result that is increases or decreases during the whole time.

The ARMA model is more complicated. When the lag is 1, the function result follows the MA(1)'s ACF. When the lag is greater than 2, the results of function are decreasing or increasing. The result of PACF follows AR(1)'s PACF when the lag is 1. As the lag increases, the result has a form like a sine curve[12][13].

To achieve the ARIMA model, many transformations are attempted in order to eliminate an irregular variance and mean. Differencing is necessary to obtain a stationary mean and log-transformation is needed to achieve a stationary variance. Most of time series data is transformed because they are not allowed to satisfy the stationary assumption. The AR, MA, and ARMA model are achieved by regular pattern data. Otherwise, The ARIMA model is achieved by irregular data. Most time series data have a tendency to show an irregular pattern[13]. Based on the ACF and PACF, it is immediately clear what model is the most appropriate model for the data. Then, the exact model can derive an exact prediction value. The ACF(5) and the PACF(6) can be calculated by SPSS software[14]. We use the SPSS software in order to save calculation time.

3 Analysis of Network Traffic

3.1 Experimental Environments and Gathering Time Series Data

The traffic monitoring system is connected to intra-network in order to collect network packets. The system shows details of network traffic enabling analyses of data gathered from the source, destination, and the total number of packets transferred at each node. The total number of packets is subject to many changes in any period of time. It is called time series data for making the model to predict the future amount of packets.

Traffic monitoring was undertaken for 1 year from July 2003 to August 2004. The number of packets was accumulated by every hour. We were unable to collect the data when the network system experienced errors or power failures, or was attacked by viruses. We classified packets except abnormal situation packets.

3.2 A Procedure for Applying Time Series Models

We use time series models to predict the number of packets. In order to predict traffic based on time series models, we assume that the models satisfy the stationary assumption. When the stationary assumption is not satisfied, the original data must be transformed such as differencing and log-transformation. If the model satisfies the assumption, then we obtain the correct prediction value. Fig. 1 shows procedures of applied time series models.

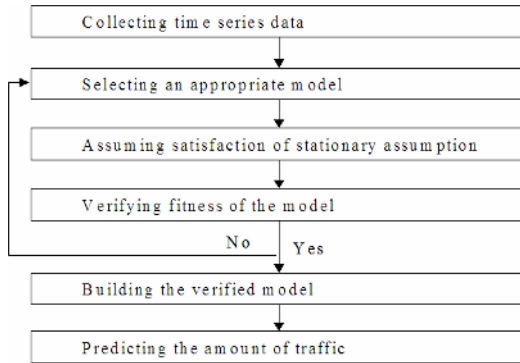


Fig. 1. Procedures of applied time series models

We obtain the results of the ACF and PACF by using SPSS whether the model satisfies the stationary assumption or not.

3.3 Assumption of Time Series Models

We collect time series data for predicting the number of packets. Traffic characteristics form a regular time series pattern over a period of time. The number of packets is subject to many changes irregularly in any period of time. Fig. 2 shows the pattern of the collected data.

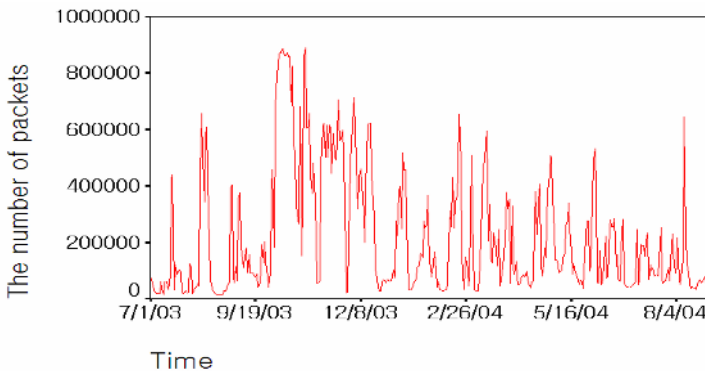


Fig. 2. The number of packets by time passing

In order to obtain amount of packets in near future, we must check the stationary assumption before applying models. The collected traffic data for 1 year are calculated into ACF and PACF. It is necessary to assume that the collected data follow a stationary process. Fig. 3 shows the results of ACF and PACF that present collected data for 1 year. Following figures are calculated by SPSS[14].

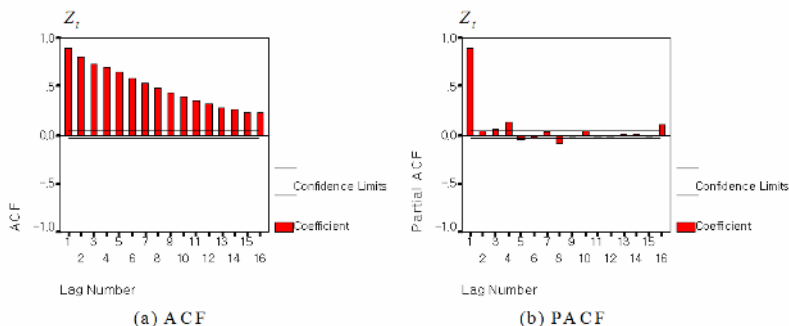


Fig. 3. ACF and PACF by yearly collection

In Fig. 3, we find that it is not possible to identify a suitable model because the results of ACF and PACF do not have any significant value which exists in the confidence intervals. The above graph indicates that the time series model does not satisfy the stationarity because the result of the function does not exist in the confidence interval at a certain lag. It explains that the correlation of original variables does not continue at present time. As a result, we conclude that original variables obtained by collecting data on a yearly basis are not suitable for predicting network traffic. In the case of non-stationary data, many transformations of the original variable are attempted to achieve a suitable time series model. In other words, transformations are used in order to eliminate an irregular mean and variance. For example, differencing and log-transformation are used in such situations. Fig. 4 shows the tendency of packets after both differencing and log-transformation.

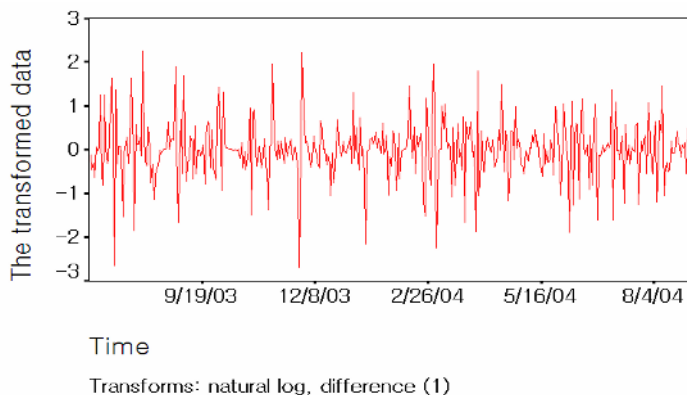


Fig. 4. The tendency of packets after both differencing and log-transformation

In Fig. 4, the tendency of transformed data is also irregular and the data do not have a constant variance. Fig. 5 shows the ACF and PACF results after the original data are transformed.

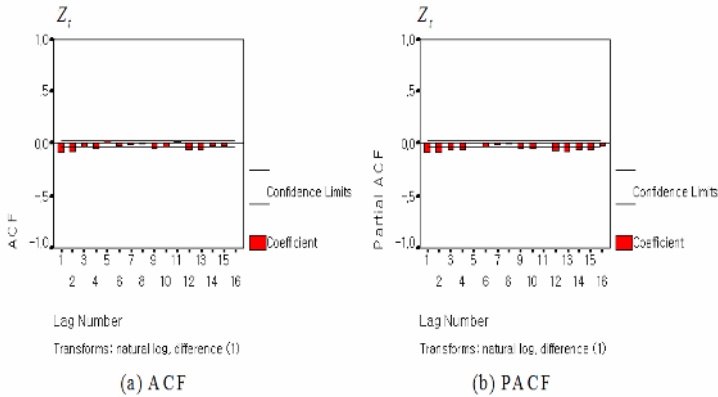


Fig. 5. ACF and PACF after both transformations

In Fig. 5, we obtain that the transformations are not sufficient to satisfy the assumption. So, the collected data on yearly basis do not explain the correlation through the data. The long period of the gathered data is not suitable for obtaining time series models. Therefore, we will find another way to obtain a proper method for predicting network traffic. One year is very long term to predict network traffic. Therefore, another way to obtain a proper method must be found. The ACF and PACF results are also obtained by classifying every 6 months. It is also not suitable for explaining the correlation of time series data.

The short period classification to use a prediction method is more available than the long period classification. Therefore, it is necessary to classify data sets in order to achieve a better result. The data set can be divided into small pieces and applied to time series models.

4 Verifying Prediction of Network Traffic by Using Classification

We realize that the short period to use a prediction method is more effective than the long period. Therefore, it is necessary to classify data sets in order to achieve a better result. The data set can be divided into small pieces and applied to the time series model.

4.1 An Analysis of Network Traffic Based on Monthly Classification

If loss of traffic happens, then data are excluded the period of month. The loss occurred when the network system experienced errors or power failures, or was attacked by viruses. In order to construct a correct model, we classified packets except

when such situations occurred. We obtain ACF and PACF results by using 11 data from July 2003 to August 2004. Data sorted on a monthly basis does not reveal any significance in ACF and PACF results. The next table shows the results of data classified on a monthly basis.

Table 1. The results of analyses by monthly classification

Model	AR	MA	ARMA	ARIMA	Inappropriate
Results	3	1	1	0	6

Although three data are satisfied with the stationary assumption, we conclude that monthly classification is not suitable for predicting network traffic. Because, three times adoptions are not sufficient to understand the predictable method.

4.2 An Analysis of Network Traffic Based on Weekly Classification

For judging an accurate analysis, we remove data at a certain week when loss of traffic happens. We obtain ACF and PACF results by using 50 data from July 2003 to August 2004.

The data sorted on a weekly basis does not reveal any significant result through the ACF and PACF. The next table shows the results of data classified on a weekly basis.

Table 2. The results of Analyses by weekly classification

Model	AR	MA	ARMA	ARIMA	Inappropriate
Results	17	4	2	0	27

17 data classified on a weekly basis follow the AR model. Otherwise, The majority of above data does not have any significance of fitness that the original data follows a time series model. It means that the data classified on weekly basis do not have any significance to predict value. The data based on weekly classification do not follow the stationary assumption.

4.3 An Analysis of Network Traffic Based on Daily Classification

We investigate original data 337 times for judging an accurate analysis. The classified data set every month and every week is not suitable into a model. But the data sorted on a daily basis reveals a significant result through the ACF and PACF. The next table shows the results of data classified on a daily basis.

Table 3. The results of analyses by daily classification

Model	AR	MA	ARMA	ARIMA	Inappropriate
Results	275	10	26	15	11

We obtain the result that the most of data classified by every day satisfies the stationary assumption. So we will use the AR model for forecasting the total number of packets in the future. Above data are modeled by using daily classifying methods. The analysis indicates that the AR(1) model has the best predictive result based on the satisfaction of stationary assumption. Ultimately, although the data set is large in size, it is necessary to divide the set in order to achieve a satisfied result. Then, the conclusion is that the AR(1) is the best model for predicting the total number of packets in the near future.

4.4 Results of the Analysis About Network Traffic

For predicting network traffic, we have executed time series models by using correlation of accumulated data. In process of prediction, data have a probable error. The larger the probable error is, the lower the degree of prediction's confidence is. In this paper, we use time series models like the AR, MA, ARMA, and ARIMA widely used in Statistics to predict network traffic. For assessing the suitability of models, using time series models is possible and it allows us to predict network traffic in near future. Verification of suitability is identified through the stationary assumption. Network traffic data on a yearly basis are not an appropriate prediction model. The transformed data also do not satisfy the stationary assumption. If data are collected over the long-term, then the collected data are not suitable for prediction. In addition, the data transformed by differencing and log-transformation also did not satisfy the stationary assumption. The most efficient method for prediction is classification of the data set on a daily basis. Thus, classification on a daily basis is an appropriate method of predicting network traffic. In order to predict network traffic, the order of AR model is 24 when the lag number is 1. Then, the prediction results are trustworthy.

The most efficient method for prediction is classification of the data set on a daily basis. In a daily classification, it is found that 81% of data in 337 trials satisfied the stationary assumption of the AR model. Thus, classification on a daily basis is an appropriate method of predicting network traffic. As a result, we conclude that classification on a daily basis is satisfied with the AR model. The AR model is the best predictable method for predicting network traffic.

5 Conclusions

Monitoring network traffic system provides the accumulated number of packets only. It gives limited information to the manager. We need more information to manage the network performance efficiently. When the predicted number of packets is provided, avoiding network congestion and efficient dynamic management can be achieved. We are looking for a routing policy applied traffic prediction results for network management. By using network traffic prediction, a new routing method and best-effort route decisions may be proposed. The proposed routing method may increase network performance. Introducing a practical routing algorithm using the traffic prediction is an important future study.

References

1. Ryan Kastner, Elaheh Bozorgzadeh and Majid Sarrafzadeh.: Predictable routing, Computer Aided Design, 2000. ICCAD-2000. IEEE/ACM International Conference (2000) 5-9
2. Yantai Shu, Zhigang Jin, Lianfang Zhang, Lei Wang: Traffic Prediction Using FARIMA Models, IEEE International Conference on Communications (1999) 891-895
3. Xun Su and Gustavo de Veciana: Predictive routing to enhance QoS for stream-based flows sharing excess bandwidth, Computer Networks, Volume 42, Issue 1 (2003) 65-80
4. William Stallings: SNMP, SNMPv2, SNMPv3, and RMON 1 and 2, Addison Wesley (1999)
5. Wilinger,W., Wilson,D., Taqqu, M.: Self-similar Traffic Modeling for Highspeed Networks, ConneXions (1994)
6. Daniel R. Figueiredo, Benyuan Liu, Anja Feldmann, Vishal Misra, Don Towsley and Walter Willinger: On TCP and self-similar traffic, Performance Evaluation, In Press, Corrected Proof (2005)
7. S. Andersson, T. Ryden: Local dependencies and Poissonification: a case study, Performance Evaluation, Volume 52, Issue 1 (2003) 41-58
8. M. Ayedemir, L. Bottomley, M. Coffin, C. Jeffries, P. Kiessler, K. Kumar, W. Ligon, J. Marin, A. Nilsson, J. McGovern et al.: Two tools for network traffic analysis, Computer Networks, Volume 36, Issues 2-3 (2001) 169-179
9. Hans-Werner Braun, Kimberly C. Claffy: Web traffic characterization: an assessment of the impact of caching documents from NCSA's web server, Computer Networks and ISDN Systems, Volume 28, Issues 1-2 (1995) 37-51
10. Wilinger,W., Wilson,D., Taqqu, M.: Self-similar Traffic Modeling for Highspeed Networks, ConneXions (1994)
11. W. Leland, et al.: On the Self-Similar Nature of Ethernet Traffic (extended version), IEEE/ACM Transactions of Networking, Vol. 2, no. 1 (1994) 1-15
12. William W. S. Wei: Time Series Analysis, Addison-Wesley (1990)
13. George E. P. Box, Gwilym M. Jenkins: Time Series Analysis: Forecasting and Control, HOLDEN-DAY (1976)
14. SPSS for windows Trends Release 11.0, SPSS Inc. (2001)

An Obstacle Avoidance Method for Chaotic Robots Using Angular Degree Limitations

Youngechul Bae¹, MalRey Lee², and Thomas M. Gatton³

¹ Division Electronic Communication and Electrical Engineering of Yosu Nat'l University,
Yosu, Chollanamdo, South Korea

² School of Electronics & Information Engineering, ChonBuk National University,
664-14, 1Ga, DeokJin-Dong, JeonJu, ChonBuk, 561-756, Korea
mrlee@chonbuk.ac.kr

³ School of Engineering and Technology, National University,
11255 North Torrey Pines Road, La Jolla, CA 92037, USA

Abstract. This paper presents a method to avoid obstacles that have unstable limit cycles in a chaos trajectory surface using angular degree limits. It is assumed that all obstacles in the chaos trajectory surface have a Van der Pol equation with an unstable limit cycle. When a chaos robot meets an obstacle in a Lorenz, Hamilton and Hyper-chaos equation trajectory that exceed the defined angular degree limits, the obstacle repulses the robot. Computer simulation of the Lorenz equation and the Hamilton and hyper-chaos equation trajectories, with one or more Van der Pol equations as the obstacle(s) is performed and the proposed method is verified through simulation of the chaotic trajectories in any plane, which avoids the obstacle when it is found, where the target is either met or within close range.

1 Introduction

Chaos theory has drawn a great deal of attention in the scientific community for almost two decades. Considerable research efforts have been performed over the past few years to export these physics and mathematics concepts into real world engineering applications. Applications of chaos are being actively investigated in such areas as chaos control [1]-[2], chaos synchronization and secure/crypto communication [3]-[7], chemistry [8], biology [9] and related robotic applications [10]. Recently, Nakamura, Y. et al [10] proposed a chaotic mobile robot which was equipped with a controller to ensure chaotic motion, and whose dynamics are represented by an Arnold equation. This investigation applied obstacles in the chaotic trajectory, but did not address chaos obstacle avoidance methods. This paper proposes a method for target searching using unstable limit cycles in the chaos trajectory surface. It is assumed that all obstacles in the chaos trajectory surface have a Van der Pol equation with an unstable limit cycle. When the method identifies the target, through arbitrary wandering in the chaos trajectories derived from Lorenz, Hamilton and hyper-chaos trajectory equations, it is applied to the chaos robots. The computer

simulations show multiple obstacle avoidance with Lorenz and Hamilton equations and hyper-chaos equations. The proposed method avoids the obstacle, within a defined range, and the results are verified to demonstrate application of the target search, with chaotic trajectories in any plane, to the mobile robot.

2 Chaotic Mobile Robot Equations

A two- wheeled mobile robot is defined as the following mobile robot mathematical model shown in Fig. 1.

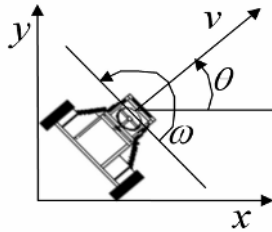


Fig. 1. Two-wheeled mobile robot

Let the linear velocity of the robot v [m/s] and angular velocity ω [rad/s] be the inputs in the system. The state equation of the two-wheeled mobile robot is written as follows:

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \end{pmatrix} = \begin{pmatrix} \cos \theta & 0 \\ \sin \theta & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v \\ \omega \end{pmatrix} \tag{1}$$

where (x,y) is the position of the robot and θ is the angle of the robot.

Chaos Equations

In order to generate chaotic motions for the mobile robot, chaos equations such as Lorenz, Hamilton and hyper-chaos equations, are applied.

Lorenz Equation

The Lorenz equation is defined as:

$$\begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= xy - bz \end{aligned} \tag{2}$$

where $\sigma = 10, r = 28, b = 8/3$. The Lorenz equation describes the famous chaotic phenomenon.

Hamilton Equation

The Hamilton equation is one of the simplest physical models and has been widely investigated through mathematical, numerical and experimental methods. The state equation of the Hamilton equation is derived as follows.

$$\begin{aligned} \dot{x}_1 &= x_1(13 - x_1^2 - y_1^2) \\ \dot{x}_2 &= 12 - x_1(13 - x_1^2 - y_1^2) \end{aligned} \tag{3}$$

Hyper-Chaos Equation

Hyper-chaos equations are also one of the simplest physical models and have been widely investigated by mathematical, numerical and experimental methods for complex chaotic dynamics. The hyper-chaotic equation is developed by using a connected N-double scroll. The state equation of N-double scroll equation is derived as follows:

$$\begin{aligned} \dot{x} &= a[y - h(x)] \\ \dot{y} &= x - y + z \\ \dot{z} &= -\beta y \end{aligned} \tag{4}$$

The hyper-chaos equation is composed from a 1 dimensional CNN (Cellular Neural Network), with two identical N-double scroll circuits. Each cell is connected by using unidirectional or diffusive coupling. This paper uses the diffusive coupling method and the state equation of x-diffusive coupling and y-diffusive coupling is represented as follows.

x-diffusive coupling

$$\begin{aligned} \dot{x}^{(j)} &= a[y^{(j)} - h(x)^{(j)}] + D_x(x^{(j-1)} - 2x^{(j)} + x^{(j+1)}) \\ \dot{y}^{(j)} &= x^{(j)} - y^{(j)} + z^{(j)} \\ \dot{z}^{(j)} &= -\beta y^{(j)}, j=1,2..L \end{aligned} \tag{5}$$

y-diffusive coupling

$$\begin{aligned} \dot{x}^{(j)} &= a[y^{(j)} - h(x)^{(j)}] \\ \dot{y}^{(j)} &= x^{(j)} - y^{(j)} + z^{(j)} + D_y(x^{(j-1)} - 2x^{(j)} + x^{(j+1)}) \\ \dot{z}^{(j)} &= -\beta y^{(j)}, j=1,2....L \end{aligned} \tag{6}$$

where, L is number of cell.

Embedding Chaos Trajectories

In order to embed the chaos equation into the mobile robot, the Lorenz, Hamilton and hyper-chaos equation are utilized as described in the following section.

Lorenz Equation

By combining equations (1) and (2), the following state variables are defined:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} \sigma(y - x) \\ \gamma x - y - xz \\ xy - bz \\ v \cos x_3 \\ v \sin x_3 \end{pmatrix} \quad (7)$$

Eq. (7) includes the Lorenz equation. The behavior of the Lorenz equation is chaotic. The chaotic mobile robot trajectory shown in Fig. 2 may be obtained by using Eq. (7) with coefficient and initial conditions as follows:

Coefficients: $v = 1$ [m/s]

Initial conditions:

$$x_1 = 0.10, \quad x_2 = 0.265, \quad x_3 = 0.27, \quad y = 0.5$$

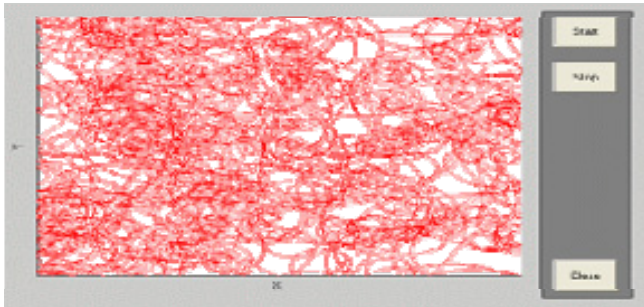


Fig. 2. Trajectory of mobile robot of Lorenz equation

3 Chaos Robot Behavior with Mirror Mapping and Obstacle Avoidance

In this section, the avoidance behavior of a chaos trajectory with obstacle mapping, relying on the Lorenz, Hamilton and hyper-chaos equation respectively, are presented.

Fig. 3 through 5 shows the chaos robot trajectories to which mirror mapping is applied in the outer wall and in the inner obstacles. The chaos robot has two fixed obstacles, and it can be confirmed that the robot adequately avoids the fixed obstacles in the Lorenz, Hamilton and Hyper-chaos robot trajectories.

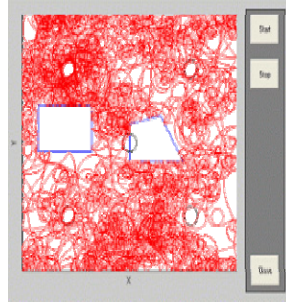


Fig. 3. Lorenz equation trajectories of chaos robot with obstacle

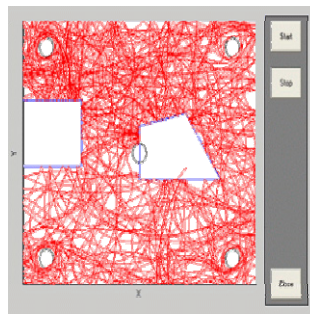


Fig. 4. Hamilton equation trajectories of chaos robot with obstacle

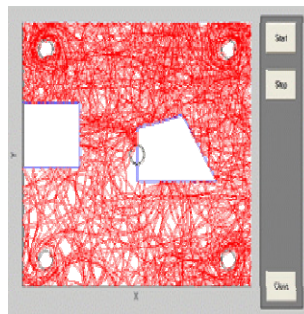


Fig. 5. Hyper-chaos equation trajectories of chaos robot with obstacles

3.1 An Obstacle Avoidance Method Using Angular Degree Limits

This section proposes a new obstacle avoidance method by using angular degree limits with Lorenz, Hamilton, hyper-chaos equations. This method ensures a safe

robot path with distance limits to avoid obstacles. This is done by constraining the limit of the angular degree when approaching obstacles.

Lorenz Equation

In Fig. 6, the robot trajectories for obstacle avoidance through angular degree limitation are demonstrated for (a) low situations and (b), high situation, respectively, in the Lorenz chaos robot.

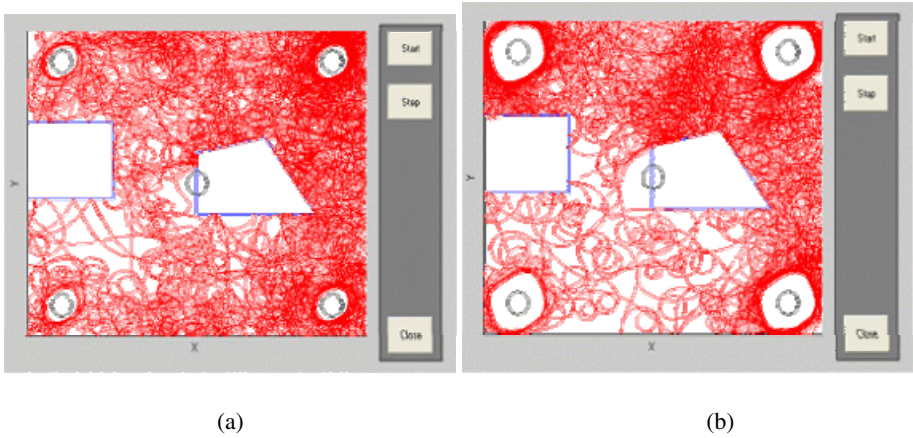


Fig. 6. Obstacle avoidance results for angular degree limits in the Lorenz chaos robot, low (a), high (b)

Hamilton Equation

Fig. 7 shows the robot trajectories for obstacle avoidance using angular degree limits for (a) low angular degree limits and (b) high angular degree limits in the Hamilton chaos robot, respectively.

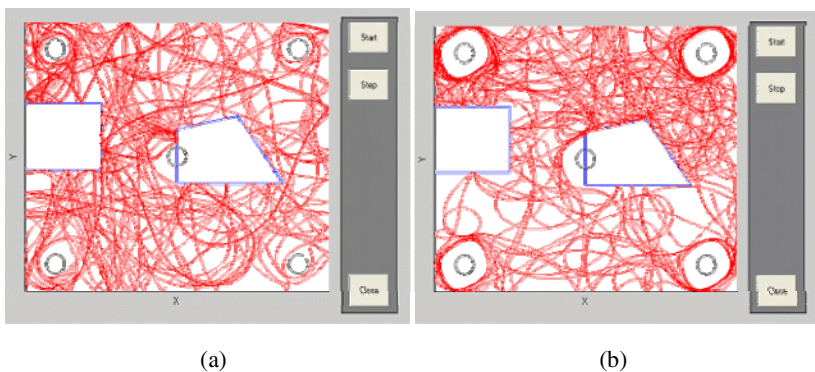


Fig. 7. An obstacle avoidance result for angular degree limits in the Hamilton equation, low (a), high (b)

4 Conclusion

This paper proposed a chaotic mobile robot, which employed Lorenz, Hamilton and hyper-chaos equation trajectories, and an obstacle avoidance method in which we assume that the obstacle has a Van der Pol equation with an unstable limit cycle.

Robot trajectories were generated representing the dynamics of mobile robots with Lorenz, Hamilton and hyper-chaos equations and integrating the proposed obstacle avoidance method using angular degree limitations. Computer simulations demonstrated that this method produces successful trajectories through the analysis of angular degree limitations.

References

1. E. Ott, C.Grebogi, and J.A York," Controlling Chaos", Phys. Rev.Lett. Vol. 64, (1990) 1196-1199
2. T. Shinbrot, C.Grebogi, E.Ott, and J.A.Yorke, " Using small perturbations to control chaos", Nature, vol. 363, (1993) 411-417
3. M. Itoh, H. Murakami and L. O. Chua, "Communication System Via Chaotic Modulations" IEICE. Trans. Fundamentals. vol.E77-A, no. (1994) 1000-1005
4. L. O. Chua, M. Itoh, L. Kocarev, and K. Eckert, "Chaos Synchronization in Chua's Circuit" J. Circuit. Systems and computers, vol. 3, no. 1, (1993) 93-108
5. M. Itoh, K. Komeyama, A. Ikeda and L. O. Chua, " Chaos Synchronization in Coupled Chua Circuits", IEICE. NLP. 92-51. (1992) 33-40
6. K. M. Short, " Unmasking a modulated chaotic communications scheme", Int. J. Bifurcation and Chaos, vol. 6, no. 2, (1996) 367-375
7. L. Kocarev, " Chaos-based cryptography: A brief overview," IEEE, Vol. (2001) 7-21
8. M. Bertram and A. S. Mikhailov, "Pattern formation on the edge of chaos: Mathematical modeling of CO oxidation on a Pt(110) surface under global delayed feedback", Phys. Rev. E 67, (2003) 136-208
9. K. Krantz, f. H. Yousse, R.W , Newcomb, "Medical usage of an expert system for recognizing chaos", Engineering in Medicine and Biology Society, 1988. Proceedings of the Annual International Conference of the IEEE, 4-7, (1988) 1303 -1304
10. Nakamura, A., Sekiguchi. " The chaotic mobile robot", , IEEE Transactions on Robotics and Automation , Volume: 17 Issue: 6 , (2001) 898 -904
11. H. Okamoto and H. Fujii, Nonlinear Dynamics, Iwanami Lectures of Applied Mathematics, Iwanami, Tokyo, Vol. 14, (1995)
12. Y. Bae, J. Kim, Y.Kim, " The obstacle collision avoidance methods in the chaotic mobile robot", 2003 ISIS, (2003) 591-594
13. Y. Bae, J. Kim, Y.Kim, " Chaotic behavior analysis in the mobile robot: the case of Chua's equation" , Proceeding of KFIS Fall Conference 2003, vol. 13, no. 2, (2003) 5-8
14. Y. Bae, J. Kim, Y.Kim, " Chaotic behavior analysis in the mobile robot: the case of Arnold equation" , Proceeding of KFIS Fall Conference 2003, vol. 13, no. 2 (2003) 110-113
15. Y. C. Bae, J.W. Kim, Y.I, Kim, Chaotic Behaviour Analysis in the Mobile of Embedding some Chaotic Equation with Obstacle", J.ournal of Fuzzy Logic and Intelligent Systems, vol. 13, no.6, (2003) 729-736
16. Y. C. Bae, J. W. Kim, Y.I, Kim, Obstacle Avoidance Methods in the Chaotic Mobile Robot with Integrated some Chaotic Equation", International Journal of Fuzzy Logic and Intelligent System, vol. 3, no. 2. (2003) 206-214

Intersection Simulation System Based on Traffic Flow Control Framework

Chang-Sun Shin¹, Dong-In Ahn², Hyun Yoe¹, and Su-Chong Joo²

¹ School of Information and Communication Engineering, Sunchon National University, Korea
{csshin, yhyun}@sunchon.ac.kr

² School of Electrical, Electronic and Information Engineering, Wonkwang University, Korea
{ahndong, scjoo}@wonkwang.ac.kr

Abstract. This paper proposes the Intersection Simulation System which can dynamically manage the real-time traffic flow for each section from the intersections by referring to the traffic information database. This system consists of the hierarchical 3 layers. The lower layer is the physical layer where the traffic information is acquired on an actual road. The middle layer, which is the core of this system, lays the Traffic Flow Control Framework designed by extending the distributed object group framework that we developed before. This layer supports the grouping of intersection, the collection of real-time traffic flow information, and the remote monitoring and control by using the traffic information of the lower layer. In upper layer, the intersection simulator applications controlling the traffic flow by grouping the intersections exist. The key idea of our study is the grouping concept that can manage the physical world in the defined logical space. The intersection simulation system considers each intersection on road as an application group, and can apply the control models of traffic flow by the current traffic status, dynamically. For constructing this system, we defined the system architecture and the interaction of components on the traffic flow control framework, and designed the application simulator and the user interface for monitoring and controlling of traffic flow.

1 Introduction

Recently, with the advancement in information and communication technologies, the Intelligent Transport System (ITS) that collects, manages, and provides real-time traffic information by using location of cars are constructed. In this system, the alleviation of traffic congestion and the traffic information service are important factors. By reflecting above requirements, the Telematics services using the traffic information service and the Location-Based Service (LBS) are provided via Internet and mobile devices [1, 2]. For constructing the ITS providing above services, we need the traffic information collecting technology, the information processing technology, the wireless network technology, the Car Navigation System (CNS) technology, the Dedicated Short Range Communications (DSRC) technology, the sensor and automatic control technology, and the traffic broadcasting technology. In addition, for providing user with the traffic information, these technologies have to be converged

and integrated with element technologies of the Geographic Information System (GIS), the Global Positioning System (GPS), LBS, and Telematics [3]. In constructing the ITS infrastructure, though some researches have defined the standard of traffic information with various wire/wireless information devices and sensors, the studies for interfaces and detail specification for the ITS applications are insufficient. Also, some of the ITS have problems to solve in inter-operating due to the heterogeneous operating system [4]. In the viewpoint of implementing technology of the ITS, the signal controlling system has been operated in congested section by monitoring the traffic condition of cars collected from the traffic information devices. But this system passively transfers the control information to remote controller according to the setting of system's administrator by using the user interface with the simple control options. Therefore, it is difficult not only to analyze the traffic information, but also to perform the dynamic controlling services for the traffic flow [5].

In this paper, we propose the Intersection Simulation System that can interoperate between the traffic information systems mentioned above and control the traffic flow dynamically. This system can construct the real-time traffic information database by collecting from the traffic information devices installed in road network to the heterogeneous ITSs, and monitor and control the real-time traffic flow on road network. Our system also provides users with the monitoring service of the real-time traffic flow via web or mobile devices, and with the controlling service of real-time traffic conditions for the traffic control center. We construct the distributed system environment of traffic information by extending the functions of the Distributed Object Group Framework (DOGF) that we have studied.

In Section 2, we describe the related works. Section 3 explains the architecture of the Intersection Simulation System, the meaning of intersection groups, and the construction method of real-time traffic information database. In Section 4, we implement the traffic flow monitoring and controlling simulator, and show various interfaces with this simulator for users. We conclude in Section 5 with summaries and future works.

2 Related Works

2.1 Traffic Information Monitoring System

We classify the information collecting system to collect and monitor traffic data as presented in the 3 kinds of systems. The first is the spot collecting system that uses loops, ultrasonic, and so on installed on the road. This system acquires the traffic volume and the section speed at a spot. The second system is the section collecting system to collect the pass time and the travel speed using beacons and probe car between sections. Finally, the third system is the qualitative collecting system using CCTV or reporters that collect the traffic jam and incident information on the road [5]. The spot collecting system cannot provide the section travelling time due to not reflecting the section characteristics. While the section collecting system can understand not only the section traveling time but also the location of cars, this system has disadvantages that require high equipment and communication cost comparing with the spot collecting system. The qualitative collecting system has an advantage

that directly obtains the traffic information with reliability. However, this system can collect the subjective traffic information by using collecting methods. Current traffic information system makes database with traffic information collected by monitoring and processes as simple format, and then serves the users. The representative media are the Internet, the radio traffic broadcasting, and the Variable Message Sign (VMS). Recently, PDA and CNS are widely used. Up to now, the real-time information collecting systems in ITS support services providing the simple traffic information. In the future, this system must support the monitoring of dynamic road status and the intelligent real-time control by using the collected information.

2.2 Distributed Object Group Framework

We have been studying the Distributed Object Group Framework for constructing the group management of server objects executing the distributed application service and the logical single system environment [6, 7, 8]. Our framework provides the distributed transparency for complicated interfaces among distributed objects existing in physical distributed system, and locates between Commercial-Off-The-Shelf (COTS) middleware tier and distributed application tier. The distributed applications located on the framework's upper layer are not mission critical applications. That is, the distributed applications support non real-time or real-time applications by using the acquired data from sensors and devices or information systems according to the property of application services. Figure 1 is showing the architecture of the DOGF.

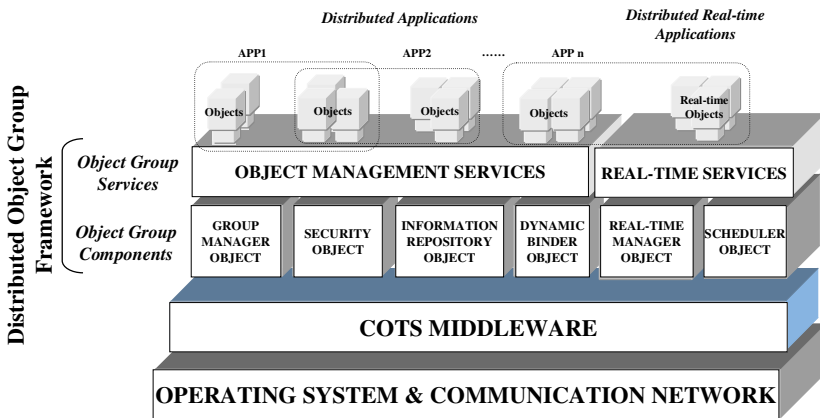


Fig. 1. Architecture of Distributed Object Group Framework

Let's show the detailed components of the DOGF. For supporting the group management service for the server objects, our framework includes the Group Manager (GM) object, the Security object, the Information Repository object, and the Dynamic Binder object with server objects. For real-time services, the Real-Time Manager (RTM) objects and Scheduler objects exist in our framework.

The key idea of our study is the grouping concept. We design the real world as the application space on distributed system. This paper considers the intersection as an application service group by extending the DOGF, and proposes the Intersection Simulation System that can control traffic flow to the real-time traffic status by configuring the intersection network via inter-communication of intersection groups and information collected from monitoring devices for traffic flow.

3 Intersection Simulation System

The road condition with traffic is changed continuously. At this time, we must develop the technology which monitors and controls real-time traffic flow by using continuous or discrete traffic data collected from road or roadside. By following the above necessity, we propose the Intersection Simulation System. Traffic flow control in intersection influences whole road condition. Therefore, if we manage the road efficiently by configuring road network interconnecting the intersections, we can guarantee the more efficient traffic flow.

3.1 System Architecture

The Intersection Simulation System must have an adaptable structure for traffic flow controlling technologies. Application layer grouping intersections exists at upper layer in our system. The middle layer of second part is the framework layer. This layer includes the road and the traffic information extractor for providing the components in the traffic flow control framework and the application objects in upper layer with physical data obtained from the traffic information collecting devices. Let's explain the functions of components of the framework layer in detail. There exists the Intersection Group Manager. This module manages the intersection group and is responsible for the registration of new application group when extending road network. The Real-Time Traffic DB Repository manages the information obtained from the various traffic information devices as traffic database. The User Interface Provider includes the simulation interface for controlling the traffic flow. The Traffic Control Model Generator exists in the framework. We can adapt the dynamic traffic flow control model to this module. The Intersection Group Interconnector is responsible for the communication among intersections. Lower layer is a physical layer as an infrastructure. In this layer, we support devices collecting the traffic information in real road and their communication. Figure 2 shows the Intersection Simulation System environment.

The execution procedures of our system are as follows. First, the system collects the traffic information through the LBS & ITS Infrastructure from the devices for the traffic information collecting on the road. Then the system extracts the effective traffic information in the system from the collected information and stores this information into the real-time traffic information database. In the intersection application simulator, to control traffic flow, we can adapt the traffic information stored by using the User Interface Provider and the traffic model provided by the Traffic Control Model Generator.

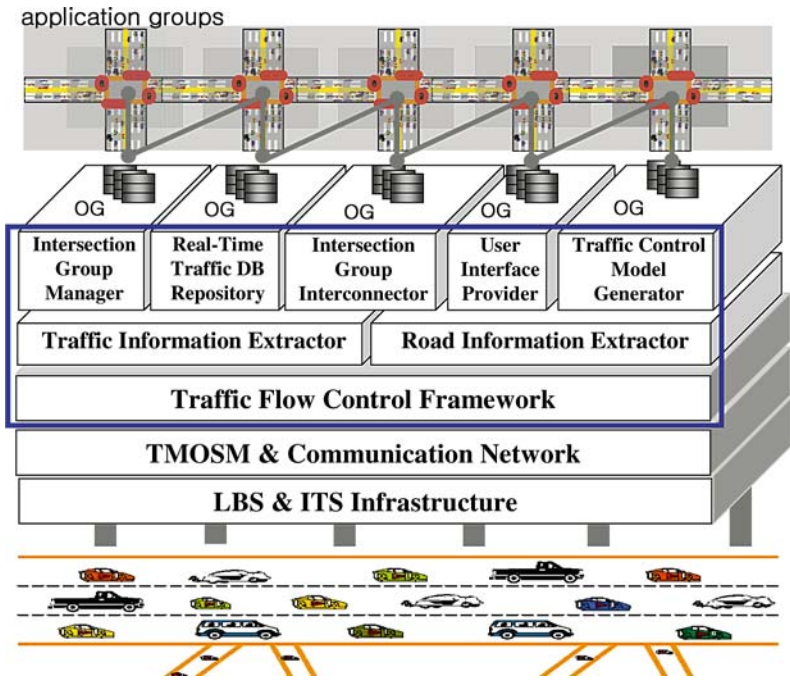


Fig. 2. Intersection Simulation System Environment

3.2 Intersection Network Group

The road consists of intersections. That is, when we connect the neighboring intersections, the road is complete. At this time, if we control the traffic flow of current intersection by referencing to the traffic information of the passed intersection according to your direction, we can provide more efficient traffic flow. Our research is based on the intersection network group. This group defines intersections as logical

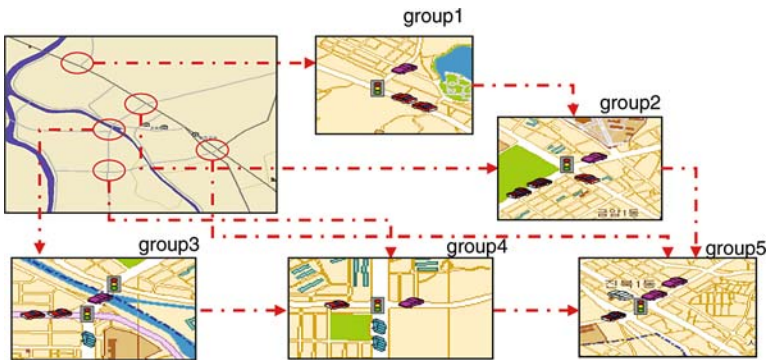


Fig. 3. Intersection Network Groups

single group and can manage road network through information exchange and control among groups. Figure 3 shows the intersection network group we mentioned.

The intersection network group makes traffic information by connecting whole road according to providing traffic flow via the group communication among intersections. We manage the status such as the opening of new road or road repairing flexibly by using the Intersection Group Manager of components of framework in Intersection Simulation System based on the intersection network group.

3.3 Real-Time Traffic Information Database

The traffic information stored in the real-time traffic information database of our system are collected by operating together real-time traffic information with the ITS and the LBS server. As shown in Figure 4, the traffic information provided by ITSs is stored into database following the standard of ITS center in the LBS server.

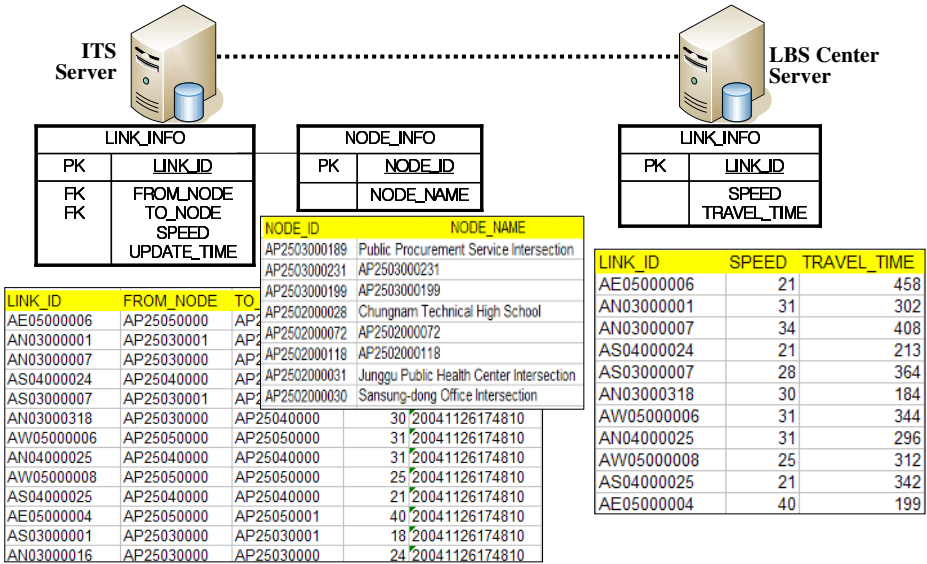


Fig. 4. Real-Time Traffic Information Database of LBS & ITS

The traffic information used in this paper is provided by ITS. This information is collected by loop installed in real road, and our system only uses the travel speed and time of section.

4 Intersection Application Simulator

In this Section, we implement the intersection application simulator that can show the traffic flow of road and intersections conveniently by using real-time traffic information database. In this simulator, we can monitor the real-time intersection traffic and control traffic flow.

4.1 Application Simulator Based on Intersection Group

The functional goals of the application simulator in the Intersection Simulation System are as follows. First, we can alleviate the traffic congestion in intersection. Second, we can preserve the easy traffic flow among intersections. To support the above functions, our system monitors car traffic and controls the signal of intersection. For implementing the simulator, we use the Time-triggered Message-triggered Object (TMO) scheme and the TMO Support Middleware developed by DREAM Lab. at University of California at Irvine. The TMO has the Service Method (SvM) triggered by client's request and the Spontaneous Method (SpM) that can be spontaneously triggered by the defined time in an object by extending the executing characteristics of existing object. This object interacts with others by remote calling [9, 10]. In our system, the objects monitoring and controlling the traffic flow in each road are implemented by TMO scheme, and we manage the TMOs as an intersection group. Now we explain the structure and functions of the TMOs. The E_TMO, the east road TMO, manages the traffic information of the right lane of road in intersection simulation environment and displays in the GUI by sending the moving information of cars to the Monitor_TMO. In the same manner, the W_TMO, the

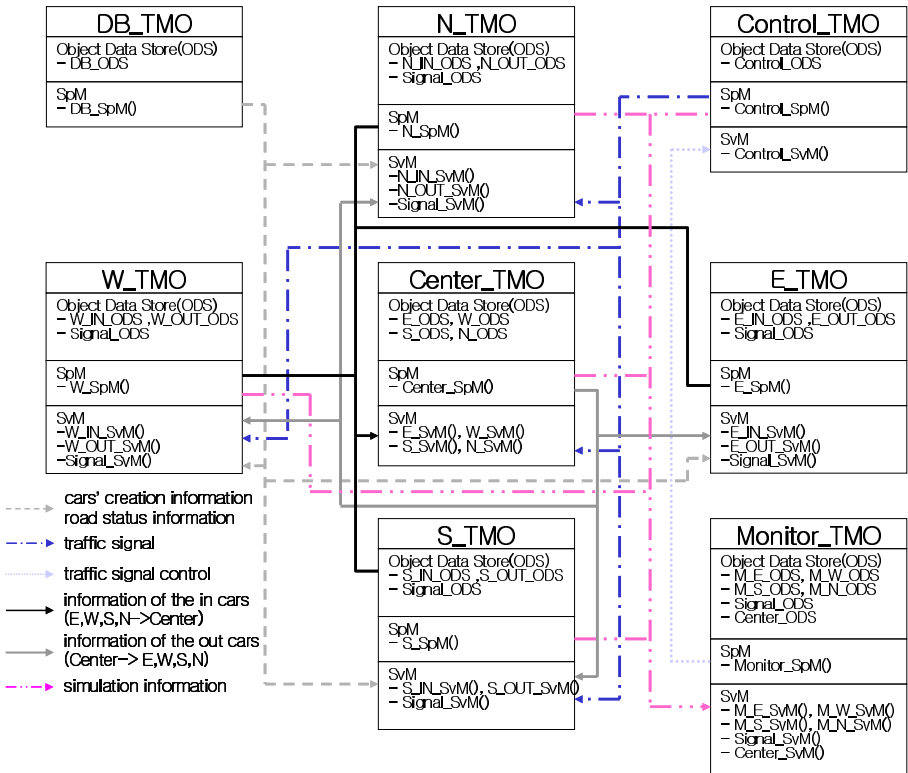


Fig. 5. Structure and Interactions of Application TMOs

S_TMO, and the N_TMO also have the identical executing structure. The Center_TMO controls the traffic information of intersection's center area and interacts with each road TMO. The Monitor_TMO transfers the all traffic information to the GUI. The DB_TMO sends the information of cars to each road TMO. Here, the information is the moving cars from the neighbor intersection group to the current intersection group according to the access direction. The Control_TMO is responsible for the signal control of intersection. The intersection system administrator monitors the traffic condition by using the GUI. And when the road is heavily congested, in GUI, we manage the traffic flow of cars with better traffic condition by controlling the signal according to the traffic control model adapted in Intersection Simulation System via the Control_TMO. Figure 5 describes the structure and the interacting procedures of TMOs configuring the intersection application simulator.

4.2 User Interfaces

The user interface collects the real-time traffic information in road by using the TMOs through individual communication network, and provides users with the increasing rate of cars, the average speed, and the signal's on/off status in each section. In Figure 6, the GUI is reflecting the traffic information transferred to the Monitor_TMO by each section at an intersection.

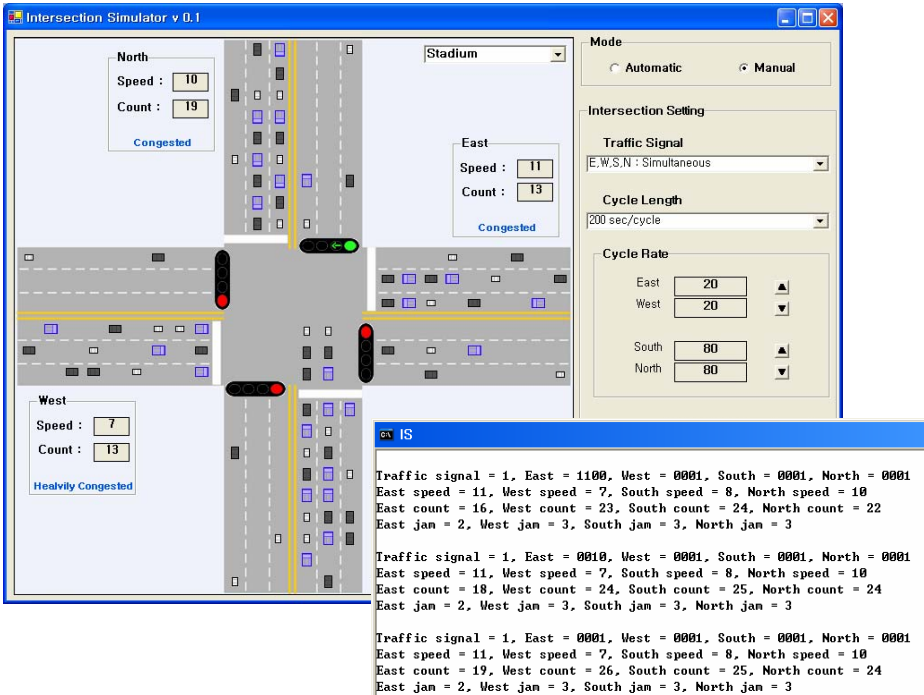


Fig. 6. User Interface of Intersection Application Simulator

For improving the traffic flow of road, user controls the on/off signal of each section by using the control components for a signal lamp based on road condition provided by the GUI. As we mentioned above, we verified that the user interface of the application simulator could monitor the real-time traffic flow in intersections and control the intersection condition according to the user requirements dynamically.

5 Conclusions and Future Works

In this research, we constructed the traffic information database by using the acquired data from the traffic information devices installed in road network, and, proposed the intelligent Intersection Simulation System which can dynamically manage the real-time traffic flow for each section of road from the intersections. We also implemented the intersection application simulator that can remotely monitor and control the traffic flow to real-time traffic condition of cars by grouping intersections. This system provides users with traffic information and services as components through the database of traffic information transferred from the ITS. And, with above way, we can execute the load balancing of servers. Hence our system can solve the problems of monitoring service of current centralized traffic condition. And if we would adapt various control models and methods, this system will alleviate traffic congestion and operate the ITS more efficiently.

In the future, we are to extend the proposed system as an intelligent system which can monitor more various traffic information provided from the ITS and control the traffic flow. And then, we will verify the excellency of our system by comparing and analyzing with the existing systems.

Acknowledgements. This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment)(IITA-2005-C1090-0501-0022).

References

1. Guk-Hyun Yoo: ITS Strategies of the Ministry of Information and Communication of Korea. Journal of the Korea ITS Society, Vol. 3, No.1 (2001) 63-73
2. Yeon-Jun Choi, Min-Jeong Kim, Moon-Soo Lee, O-Cheon Kwon: The Way of Design of a Telematics Software Middleware. In Proceedings of the Korea Information Processing Society, Vol.11, No.2 (2004) 1667-1670
3. Young-Jun Moon: Telematics Industry Activation and Traffic information Service Methods. Journal of the Korea Information Processing Society, Vol.11, No.4 (2004) 63-68
4. Shunsuke Kamijo, Yasuyuki Matsushita, Katsushi Ikeuchi, Masao SakauchiKim, K.H., Ishida: Traffic Monitoring and Accident Detection at Intersections. In Proceedings of the International Conference on Intelligent Transportation Systems (1999) 703-708
5. Jeung-Ho Kim: The Present and Vision of BEACON System. Journal of the Korea ITS Society, Vol.2, No.1 (2004) 30-39
6. Chang-Sun Shin, Chang-Won Jeong, Su-Chong Joo: TMO-Based Object Group Framework for Supporting Distributed Object Management and Real-Time Services. Lecture Notes in Computer Science, Vol.2834. Springer-Verlag, Berlin Heidelberg New York (2003) 525-535

7. Chang-Sun Shin, Su-Chong Joo, Young-Sik Jeong: A TMO-based Object Group Model to Structuring Replicated Real-Time Objects for Distributed Real-Time Applications. *Lecture Notes in Computer Science*, Vol.3033. Springer-Verlag, Berlin Heidelberg New York (2003) 918-926
8. Chang-Sun Shin, Chang-Won Jeong, Su-Chong Joo: Construction of Distributed Object Group Framework and Its Execution Analysis Using Distributed Application Simulation. *Lecture Notes in Computer Science*, Vol.3207. Springer-Verlag, Berlin Heidelberg New York (2004) 724-733
9. Kim, K.H., Ishida, M., Liu, J.: An Efficient Middleware Architecture Supporting Time-triggered Message-triggered Objects and an NT-based Implementation. In *Proceedings of the IEEE CS 2nd International Symposium on Object-oriented Real-time distributed Computing (ISORC'99)* (1999) 54-63
10. K.H(Kane). Kim, Juqiang Liu, Masaki Ishida: Distributed Object-Oriented Real-Time Simulation of Ground Transportation Networks with the TMO Structuring Scheme. In *Proceedings of the IEEE CS 23rd International Computer Software & Applications Conference (COMPSAC'99)* (1999) 130-138

A HIICA(Highly-Improved Intra CA) Design for M-Commerce*

Byung-kwan Lee¹, Chang-min Kim², Dae-won Shin³, and Seung-hae Yang⁴

^{1,3,4} Dept. of Computer Engineering, Kwandong Univ., Korea
bklee@kd.ac.kr, sdw1951@hanmail.net, yang7177@chollian.net
² Dept. of Computer Science, Sungkyul Univ., Korea
kimcm@sungkyul.ac.kr

Abstract. HIICA(Highly-improved Intra CA) proposed in this paper is internal CA which highly improved by using OCSP(Online Certificate Status Protocol)and LDAP(Lightweight Directory Access Protocol). It provides more security and confidentiality than external CA. As HIICA is provided with certification policy from PCA(Policy Certificate Authority) and sets an special environment like subscriber registration process, it reduces additional expense and strengthens the security at the certificate request process through HIICA agent. Therefore, secure electronic payment system is optimized through HIICA design that removed the existing CA complex process and unsecure elements.

1 Introduction

Fig.1 shows the structure of HIICA that simplifies external CA and which is different from existing complex external CA. HIICA which is provided with certification policy from higher CA, or PCA(Policy Certificate Authority) issues certificate to the other client requests, renews certificate and revokes certificates. Therefore, this simplified HIICA decreases the maintenance cost and the other cost that clients must pay. In addition, this paper proposes an optimized validity test process using OCSP(Online Certificate Status Protocol) and LDAP(Lightweight Directory Access Protocol), which reduces the downloading time and communication traffic.

2 Related Work

WAP(Wireless Application Protocol) is an open international standard for applications that use wireless communication, for example, internet access from a mobile phone.

WAP was designed to provide services equivalent to a Web browser with some mobile-specific additions, being specifically designed to address the limitations of very small portable devices. It is now the protocol used for the majority of the world's mobile internet.

* This research was supported by the program for the Training of Graduate Students in Regional Innovation, which was conducted by the Ministry of Commerce, Industry and Energy of the Korean Government.

WTLS(Wireless Transport Layer Security) provides a public-key cryptography-based security mechanism similar to TLS(Transport Lay Security). Also, transport-level security protocol is based on the Internet security protocol known as TLS. WTLS can authenticate communicating parties, encrypt and check the integrity of the data when it is in transit. WTLS has been optimized for use in wireless devices that rely on narrow bandwidth wireless networks. WTLS is a cryptography-based, PKI-enabled protocol that provides the security to WAP applications.

2.1 WPKI

Wireless Application Protocol PKI(WPKI) is not an entirely new set of standards for PKI; it is an optimized extension of traditional PKI for the wireless environment. WPKIs, like all PKIs, enforce m-commerce business policies by managing relationships, keys and certificates. WPKI is concerned primarily with the policies that are used to manage E-Business and security services provided by WTLS and WMLS Crypt in the wireless application environment. In the case of wired networks, IETF PKI standards are the most commonly used; for wireless networks, WAP Forum WPKI standards are the most commonly used. The WPKI is an networked server, like the WAP Gateway, it logically functions as the RA and is responsible for translating requests made by the WAP client to the RA and CA in the PKI. The PKI Portal will typically embed the RA functions and interoperate with the WAP devices on the wireless network and the CA on the wired network.

2.2 Certification Structure

Certification PKI provides technical and operational foundation for proving subscriber's identity safely and it's core is security function. Because of using authentication issued by public certification authority rather than password based authentication in electronic transaction under cyber space, WPKI(Wireless Public Key Infrastructure) has strong security function. SSL is a commonly-used protocol for managing the security of a message transmission on the internet.

3 HIICA(Highly-Improved IntCA) Design

HIICA has the effect of interference elimination by removing the mutual authentication process of external CA. The advantage of HIICA in authentication procedure is as follows. First, after environment is set and program is run once, there is no additional expense. Second, it reduces processing time and cost by simplifying certification process between the transaction persons concerned. Third, compared to external CA, HIICA strengthens security by dually securing certificates request. Forth, processing time and procedure that takes to update certificate are shortened. Fifth, because HIICA which issues certificate gives a subscriber a certificate easily through program, it performs certification function easily.

3.1 Design of HIICA

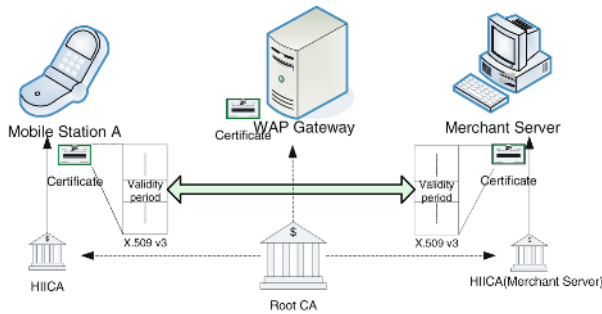


Fig. 1. HIICA Structure

3.1.1 HIICA System

Fig.2 shows the structure of HIICA that is different from existing external CA. HIICA consists of wired/wireless CA, X.509 HIICA Server, IntRA Server and WTLS HIICA Server. In the WPKI Server and Web Server, HIICA performs registration process through each server.

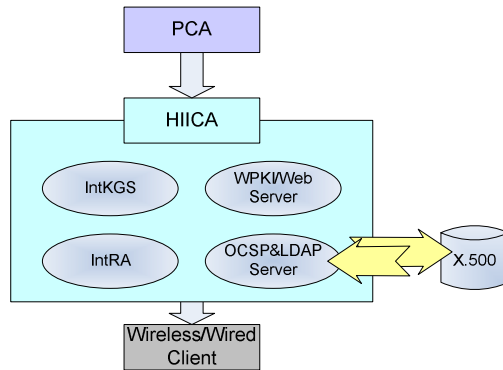


Fig. 2. The elements of HIICA architecture

The following steps are the process for the request of HIICA to be entrusted with an authorization from PCA which is the upper level.

step1] PCA generates a random nonce and transfers it to HIICA which requests certification authorization. step2] HIICA generates random number and VID(Virtual Identifier) with hash function. step3] The nonce which was transferred from PCA is made into message digest with hash function. step4] The message concatenated with certificate authorization request and VID and the H(nonce) are encrypted and is transferred to PCA. step5] PCA decrypts the transferred ciphertext (M||H(nonce)). step6] The message digest made into the nonce-1 with hash function and the

decrypted $H(\text{nonce})$ are compared and verified for integrity. step7] As integrity check is completed, PCA provides certification authorization to HIICA.

3.1.2 Design of IntKGS

When a subscriber applies for certification, IntKGS(Int Key Generation System) generates key pair and registers the keys in the certificate, or when his key by intruder is exposed, IntKGS(Int Key Generation System) regenerates the key and reregisters the keys pair in the certificate. That is, IntKGS generates a public key and a private key, registers the public key in user's certificate using LDAP(Light Directory Access Protocol), and opens the certificate to the party concerned. IntKGS selects an irreducible polynomial $f(x)$ and elliptic curve E , and decides vector a and b of E .

Fig. 3 shows the F2mECC algorithm of key generation in order to register the values($f(x)$, E , a , b , p and k with common list.

<pre>ecc_key_generation(f(x), a, b, E, P) input_private_key(k); for (i=k ; i>=1 ; i--) if (x1 == x2 && y1 == y2) L = x1 +y1 * Multiply_inverse(x1); x3 = L2 + L + a; y3 = x12 + (L + 1) * x3; x2 = x3; y2 = y3;</pre>	<pre>else if (x1 != x2 && y1 != y2) L = (y2 + y1) * Multiply_inverse(x2 + x1); x3 = L2 + L + x1 + x2 + a; y3 = L*(x1+x3) + x3 + y1; x2 = x3; y2 = y3; return(kP);</pre>
--	---

Fig. 3. Public key generation algorithm

3.1.3 The Design of HIICA System

1. The handshake of wired/wireless certification

Fig. 4. shows the process of certificate handshake. A sender sends a message that digitally signed a sender's header with an HIICA's private key. A receiver receives the message, and verifies the message integrity. And then the receiver requests the body part to a sender. Finally, the receiver received the message that digitally signed the sender's body part with an HIICA's private key.

<pre>intX509_CA_Body_request(char *encrypt_header) { char *digest1, *digest2, *pkey, *encrypt_body; intX509_header h; pkey = read_public_key(); read_header(); digest1 = decrypt_header(encrypt_header, pkey); digest2 = hash(h->version, h->name, h->period, h->header_sign); encrypt_body = send_body(); else{ printf("--- data integrate ---");</pre>	<pre>return(); } return(encrypt_body); } void send_body(){ intX509_body b; char *digest, *encrypt_body ; pkey = read_private_key(); read_body(); digest = hash(b->s_no, b->algo_name, b->s_name, b->p_key, b->body_sign); encrypt_body = Encrypt_sign(digest, pkey); return(encrypt_body); }</pre>
--	---

Fig. 4. The message integrity of HIICA

In the HIICA, HIICA handshakes an certificate using certificate version, issuer and validity period. If validity period is possible, HIICA receives body part. And Then, HIICA encrypts user’s messages using HIICA’s public key and sends the messages.

2. wired/wireless certification

Fig. 5 shows a generation algorithm of certificate. When a user requests a user’s certificate to HIICA, HIICA issue user’s certificate in the form of HIICA X509v3 header and body.

<pre>intX509_CA_create(char *s_name, char *name) int no; char *digest; intX509_header h; intX509_body b; FILE *fp; pkey = read_private_key(); no = read_s_no(); b->algo_name = choose(); h->version = 1; h->name = name; h->period = date(); digest = hash(h->version, h->name, h->period);</pre>	<pre>h->header_sign = create_sign(digest, b->algo_name, pkey); b->s_no = no++; b->p_key = create_public_key(); b->s_name = s_name digest = hash(b->s_no, b->algo_name, b->s_name, b->p_key); b->body_sign = create_sign(digest, b->algo_name, pkey); header_file_print(h, fp); body_file_print(b, fp);</pre>
--	---

Fig. 5. Generation algorithm of certificate

3. IntRA System

Shown in Fig 6, IntRA performs a user’s identity confirmation and registration to issue certificate instead of IntCA. That is, IntRA issues a user’s certificate over Internet, and registers it to IntCA. Also, it is possible for HIICA to manage the user’s information reference, modification, delete etc..

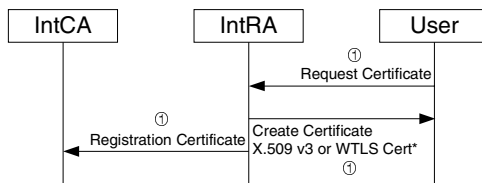


Fig. 6. IntRA Roles

3.1.4 The Display of HIICA Browser System

1. WPKI/Web server

To display the result of certificate issue on the browser to a user, HIICA uses CGI program such as Fig. 7. WPKI/Web server displays the process of HIICA such as issue the procedure of certificate and the update procedure of certificate on the user’s Web Browser.



Fig. 7. WPKI/Web server

2. OCSP & LDAP server

When a user verifies digital signature between users that have certificate, the user verifies the certificate using X.500 directory. Fig. 8 shows the request process of LDAP certificate by OCSP client's request. When clients are connected to the OCSP Server in the same time, OCSP server solves the problem using priority algorithm according to client's credit grade.

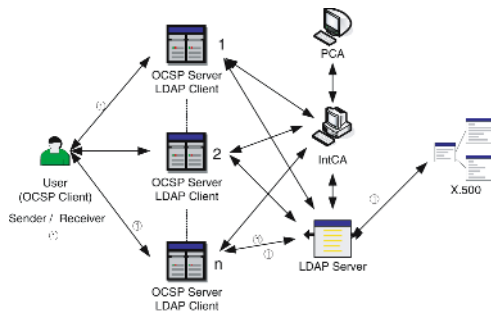


Fig. 8. OCSP & LDAP processing

This paper provides a user(OCSP Client) with certificate information by accessing LDAP server with OCSP LDAP Server, and proposes that the user can design the request and update of HIICA through OCSP server(LDAP Client).

① OCSP client application transmits unique identifier of certificate that identify status information to OCSP server. ② After OCSP Server checks and analyses users' request according to client's credit grade, it makes query and send it to LDAP server using LDAP. If the request has a fault, it is returned to OCSP client application and receives modified request again. ③ LDAP server using DAP transmits the query to X.500 Server and receives such return value as good, revoked, and unknown. ④ LDAP server transmits data which was received from X.500 Server to OCSP Server.

⑤ OCSP server identifies the received data and transmits status information of certificate to OCSP Server. ⑥ OCSP client application verifies the received data and confirms certificate status and history. Therefore, Secure electronic commerce payment protocol using HIICA is performed, after it checks the other client's certificate status through original client request.

4 Performance Evaluation

4.1 The Reduction of Certification Process

Table 1 and Fig 9. shows that certification process is reduced by verifying the validity of certificate between HIICA and external CA.

Table 1. Certification process

CA Steps	External CA	HIICA	Performance evaluation
steps	1. Client requests 2. Merchant responds 3. Merchant CRL requests to client 4. Client CA's CRL requests to Merchant Server's CA 5. Merchant Server's External CA verifies CRL 6. Merchant Server's External CA responds the result of CRL verification to client CA 7. Client CA confirms client	1. Client requests 2. Merchant responds 3. Merchant renews his own certificate by using OCSP 4. Client verifies the validity of merchant's certificate 5. Client CA confirms client	Int CA reduces 2 steps, compared to external CA.

4.2 The Comparison of ECC with RSA

In this paper, the proposed HIICA uses ECC instead of RSA. In comparison with RSA, the results of the encryption and decryption times are shown in Tables 2 respectively, which indicate that encryption and decryption time of ECC are much less than those of RSA.

Table 2. A comparison for encryption/decryption time (unit : Φ s)

Method Key size (byte)	encryption		decryption	
	RSA	ECC(F_2^m)	RSA	ECC(F_2^m)
5	0.05	0.03	0.11	0.03
10	0.54	0.03	0.55	0.04
15	1.54	0.03	1.20	0.04
20	2.55	0.04	3.08	0.04
25	4.33	0.04	6.21	0.05
50	5.53	0.04	8.06	0.04
100	7.28	0.03	9.95	0.03

4.3 The Advantage of HIICA

1. When issuing certificates, HIICA saves additional cost. 2. HIICA manages the process of certificate issue, servers and clients in the organization easily. 3. When issuing new certificates and renewing certificates, HIICA is faster than external CA. 4. The Administrator's training time for issuing certificates is reduced significantly. 5. In Fig. 5, as HIICA removes the version and the issuer name of certificate, traffic is reduced.

5 Conclusion

To perform secure mobile payment system, this proposed HIICA instead of existing external CA has the following improved performance. First, As HIICA identifies the validity period of each header and uses public key of body. HIICA protect the whole certificate from the forgery. Second, process steps is shortened, because mutual trust of internal users need not check all certificate item. Third, there is no additional cost in renewal if HIICA is established once. Fourth, the procedure is more simplified than the existing external CA. Finally, it gains reliability between transaction persons through real time certification verification that use OCSP and LDAP.

References

1. B. K. Lee, E-Commerce Security, Namdoo Books, 2002
2. N. Koblitz, "Elliptic Curve Cryptosystems", Mathematics of Computation, 48, pp.203~209, 1987.
3. V. S. Miller, "Use of elliptic curve in cryptography", Advances in Cryptology-Proceedings of Crypto'85, Lecture Notes in Computer Science, 218, pp.417~426, Springer-Verlag, 1986.
4. I. S. Cho and B. K. Lee, "ASEP Protocol Design", ICIS, Vol.2, pp.366~372, Aug. 2002.
5. S.H. Yang, Y.K. Lee and B. K. Lee, " end-to-end authentication function Using F2mECC and EC-DH", Journal of Korean Society for Internet Information, Vol. 5, No. 1, pp.403~406, 2004.
6. Y. H. Choi, "A Study on Performance Improvement of Certificate Validation Algorithm in PKI", 2003
7. S. Y. Song, "Mobile PKI Authentication Model Using Delegation Ticket, 2004.
8. <http://www.certicom.com>

Highly Reliable Synchronous Stream Cipher System for Link Encryption

HoonJae Lee

Dongseo University, Busan, 617-716, Korea
hjlee@dongseo.ac.kr

Abstract. A highly reliable synchronous stream cipher system with absolute synchronization is proposed. The proposed system¹ includes an improved initial synchronization with highly reliable keystream synchronization in a noisy channel (about BER=0.1). Furthermore, encryption algorithms with a LILI-II, a Dragon and a Parallel LM for data confidentiality, and a zero-suppression algorithm for system stabilization are all designed and analyzed.

Keywords: Keystream, stream cipher, synchronization, randomness.

1 Introduction

Symmetric cryptosystems can be classified as either block ciphers or stream ciphers. A block cipher (EBC mode) divides plaintext into blocks and then encrypts each block independently, whereas a stream cipher encrypts plaintext on a bit-by-bit (or byte-by-byte) basis. Since a stream cipher is based on an exclusive-OR (XOR) operation, the encryption/decryption of bit-stream data bit-by-bit (or byte-by-byte) using a stream cipher is very efficient. Therefore, generally, a stream cipher is much simpler and faster than a block cipher [1-3]. A block cipher is useful as regards software implementation, however, its performance is degraded in a noisy channel due to its channel error propagation properties. Conversely, a stream cipher is good for high-speed enciphering and with noisy (wireless) channels because it produces no error propagation. The major problem of a stream cipher is the difficulty in generating a long unpredictable bit pattern (keystream). In the one-time pad in a stream cipher, the keystream is a sequence of randomly generated bits, and the probability that an individual bit will be 1, independent of the other bits, is equal to one half. An ideal keystream in a one-time pad is purely random with an infinite length. Such a keystream can neither be generated by the receiving end nor distributed to the receiving end. Currently, pseudorandom bit generators are widely used to construct keystreams by generating a fixed-length pseudorandom noise. The ability to increase the length of a keystream while maintaining its randomness is crucial to the security of a stream cipher.

¹ This research was supported by University IT Research Center Project.

There are various choices for encryption endpoints. At one extreme, link encryption enciphers and deciphers a message at each node between the source node and the destination node, whereas at the other extreme, end-to-end encryption only enciphers and deciphers a message at the source and destination.

This paper investigates link encryption as a means to secure a channel between the transmitter and the receiver. A synchronous stream cipher requires the detailed design of keystream synchronization, a keystream generator [4-6], a ZS (zero-suppression) algorithm [7], and session-key distribution. Keystream synchronization matches the output binary stream of the two keystream generators at the transmitter and at the receiver. Generally, keystream synchronization can be classified into the initial synchronization and the continuous synchronization. The former only synchronizes the two keystreams initially at the transmitter and at the receiver, whereas the latter synchronizes the keystreams both initially and continuously based on a time period. The reliability of this keystream synchronization is critical to the overall system performance and communication efficiency. This paper proposes an absolute synchronization method as an improved version of the initial synchronization method. The proposed method can be applied to a noisy channel as a wireless communication link and has a highly reliable synchronization probability. Consequently, a stream cipher system is proposed which includes a secure keystream generator (a LILL-II, a Dragon and a Parallel LM for data confidentiality), a ZS (zero-suppression) algorithm for adjusting the output sequences, and a M-L key distribution protocol with authentication [8]. Finally, the keystream synchronization performance and cryptographic security of the proposed system is analyzed.

2 Keystream Synchronization and Proposed System

Stream ciphers can be classified into self-synchronous stream ciphers and synchronous stream ciphers [10]. In a self-synchronous method, keystream synchronization is established autonomously based on the feedback of the ciphertext, however, one-bit errors are propagated in the channel relative to the size of the shift register used. In contrast, keystream synchronization in a synchronous stream cipher is re-established by sync protocols when out-of-synchronization occurs. However, in this case, no bit-errors are propagated in the channel, which is why the latter cipher is more generally used [2]. In a stream cipher, the pseudo-random binary sequences must be equal at both the transmitter and the receiver. Whether they have departed or not, the output keystreams should coincide with each other, which establishes keystream synchronization. In general, keystream synchronization methods can be classified into the initial synchronization and the continuous synchronization. In the initial synchronization (fig.1-a) the two keystreams are only synchronized initially at the transmitter and at the receiver, whereas in the continuous synchronization (fig.1-b) the keystreams are synchronized both initially and periodically. Keystream synchronization exchanges the synchronization patterns (SYNPAT) at the transmitter and at the receiver to initialize the starting point of the keystream cycle. The reliability of the keystream

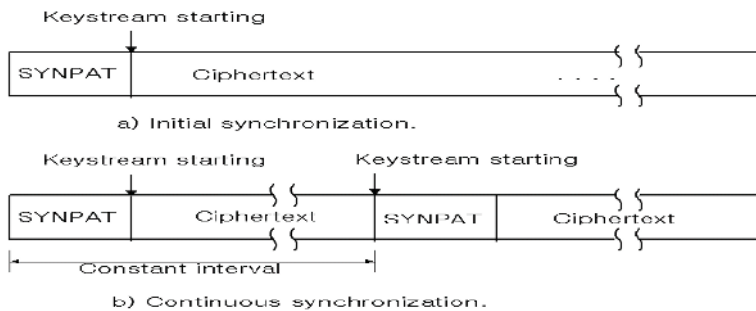


Fig. 1. Method of stream synchronization

synchronization is critical to the overall system performance and communication efficiency.

The proposed synchronous stream cipher system is shown in fig. 2 and the function of each block is as follows: Block (1) is the data terminal equipment (DTE or CODEC), block (3) is the main controller for monitoring the system and contains the master CPU, block (4) is the synchronization pattern generator that matches the keystream sequences from the sender with those from the receiver for the initial synchronization, blocks (5) and (12) are the session-key buffers that initiate the keystream generators at the sender and receiver ends, blocks (6) and (11) are the session key constructors that distribute/construct a secure session key through a public channel from the sender to the receiver, blocks (7) and (13) are the same keystream generators with high security both at the sender and the receiver ends, blocks (8) and (14) are the ZS algorithms that reduce the excessive zeros in the ciphertext at both the sender and the receiver ends, switch (9) is the data selector for the synchronization pattern, session key, or ciphertext data, block (2) is the line modem (DCE), and block (10) is the synchronization pattern detector that identifies the sync pattern in the received data.

2.1 Highly Reliable Initial Keystream Synchronization

In this paper, we propose the highly reliable initial keystream synchronization with a generator and a receiver.

A generator part and a detector part of the keystream synchronization, blocks (4) and (10) shown in fig. 2, give to match the keystream sequences at both the sender and the receiver ends. In the proposed scheme, the keystream synchronization exchanges the synchronization patterns (SYNPAT) to initialize the starting point of the two keystream cycle sequences at the sender and receiver ends. The statistical properties are selected based on the following decision criteria [9-10]. (1) A good autocorrelation property is required. (2) The same rate must be achieved on "0" and "1" in the pattern. (3) A short run should be larger than a long run.

Accordingly, the pattern is determined based on the above criterion using a Gold sequence generator [9], as in fig. 3. In this figure, the primitive polynomials

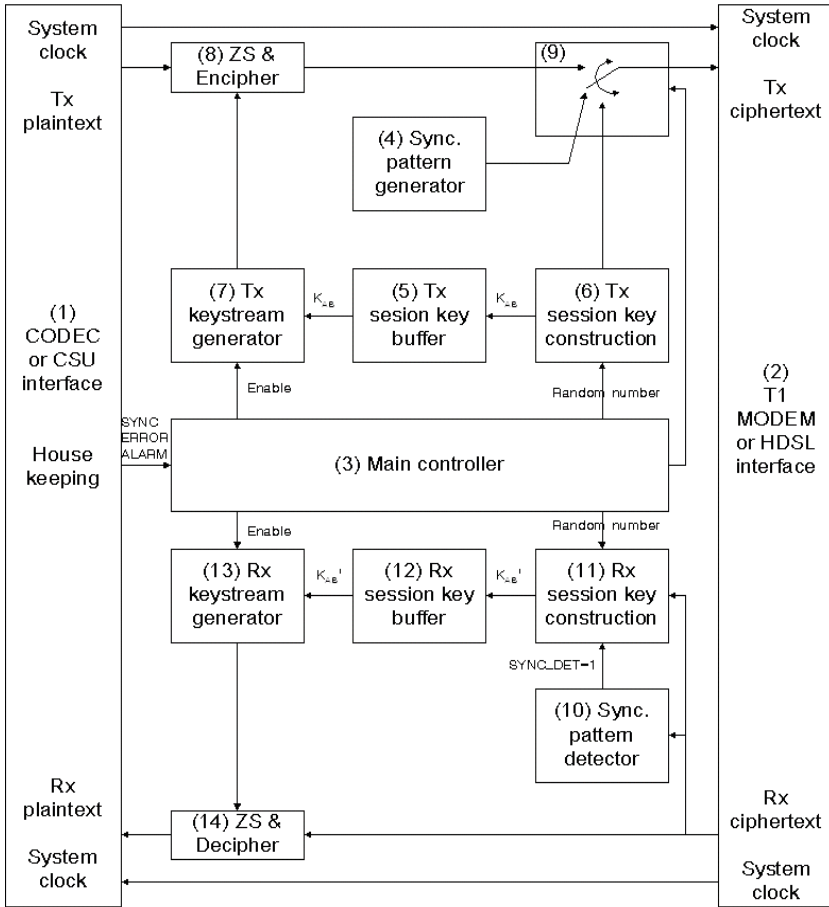


Fig. 2. Proposed synchronous stream cipher system and absolute synchronization method

of a 31-stage LFSR1 and LFSR2 were selected, then the N -bit pattern (here, $N=128$ SYNPAT) was generated and checked using the decision criteria. The two selected primitive polynomials and generated pattern are follows:

$$h_1(x) = x^{31} + x^{11} + x^2 + x + 1(31 - stageLFSR_1)$$

$$h_2(x) = x^{31} + x^9 + x^3 + x + 1(31 - stageLFSR_2)$$

$$SYNPAT = 6DDA51917C90726C7941AD046ABC8F5D(128 - bit)$$

2.1.1 Autocorrelator and Encryption Synchronization

In OFB mode of block cipher and synchronization stream cipher, cipher synchronization has transmission and reception synchronization role.

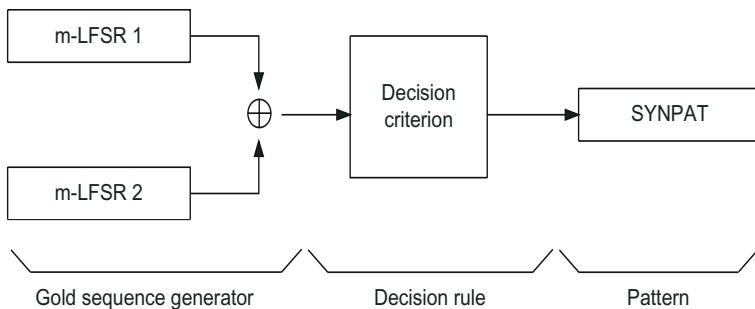


Fig. 3. Synchronization pattern generator

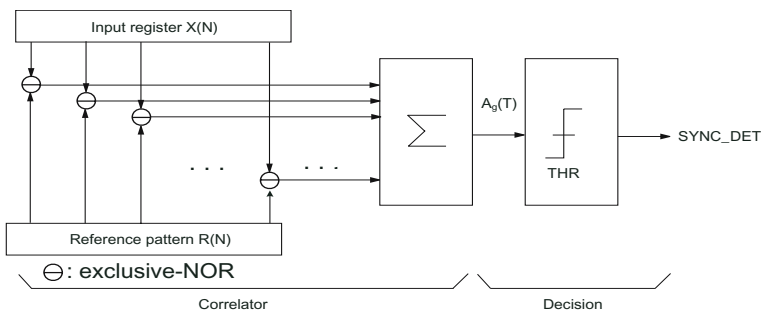


Fig. 4. Pattern detector and Pattern generator

Autocorrelation value of synchronization pattern is $A(t)$. Threshold is THR . The N_T can be found by [12].

Autocorrelation, threshold, the number of agreement and the number of disagreement are follows:

$$A(t) = \frac{A_g(t) - D_g(t)}{N}$$

$$THR = N - N_T$$

$$A_g(t) = \sum_{i=1}^N X(i) \ominus (i)$$

$$D_g(t) = \sum_{i=1}^N X(i) \oplus R(i)$$

$$A_g(t) - D_g(t) = N$$

In Fig. 4, Synchronization pattern detector have to calculate the truth and false within one clock as adjusting summation part in system clock. It is difficult to embody this in hardware.

The synchronization pattern detector at the receiver consists of a correlation part and decision part, as shown in fig. 4. The correlation part computes the number of agreements in bits, $A_g(t)$, between an input pattern and the reference pattern (*SYNPAT*). The decision part then compares the number of agreements in bits with a threshold, *THR*, and displays one of the following: If the number of agreements in bits is larger than the threshold, "synchronization detected (*SYNC - DET=1*)", otherwise "synchronization failed (*SYNC - DET=0*)".

2.1.2 Improved Autocorrelator System

When the value of N number becomes high, the synchronization pattern detector becomes more complicated, because general Synchronization pattern is formed by random value. Therefore we know that after random Synchronization pattern is changed to single pattern, and then followed by another change to Synchronization pattern, it reduces hardware complexity. In other words, after all of the bits are changed, such as repetition of all 1's pattern or 0's pattern, and 10's pattern or 1100's pattern, we can look for the number of agreement bits. It is very simple process for the hardware. Autocorrelation of single pattern itself is just low and it does not fit for Synchronization pattern so after generating better autocorrelation, we need another process to change single pattern. When transmitting the single pattern, there is a problem in wired code such as modem clock restore problems by all 1's pattern and all 0's pattern. To protect these problems, we have to add scrambler and de-scrambler to this system. Fig. 6 shows them.

In Fig. 5, the scrambler of transmission changes all bits "1" pattern to random pattern and the de-scrambler of reception shows reverse way. It could be embodied simply according of wired environment. All 1's pattern detector of reception is shown in Fig. 6. In Fig. 6, all 1's pattern or random pattern are input N-level register and "Up-Down counter" should count the number of "1". In other words, if value of "1" is put on the first place of register, the counter

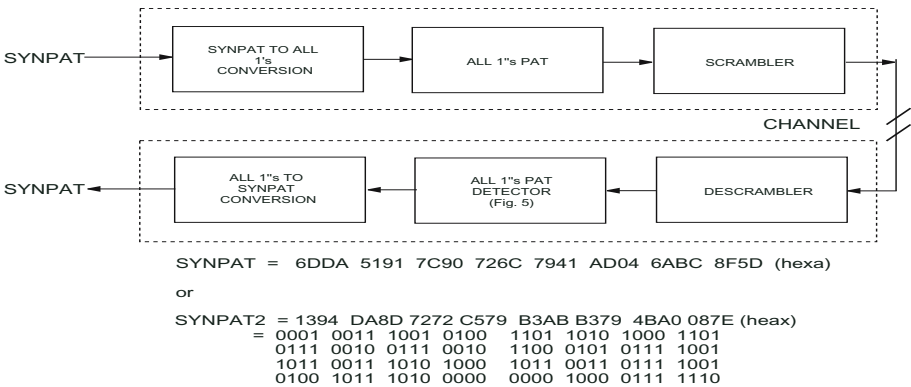


Fig. 5. Design Proposal of key random syn pattern with single pattern

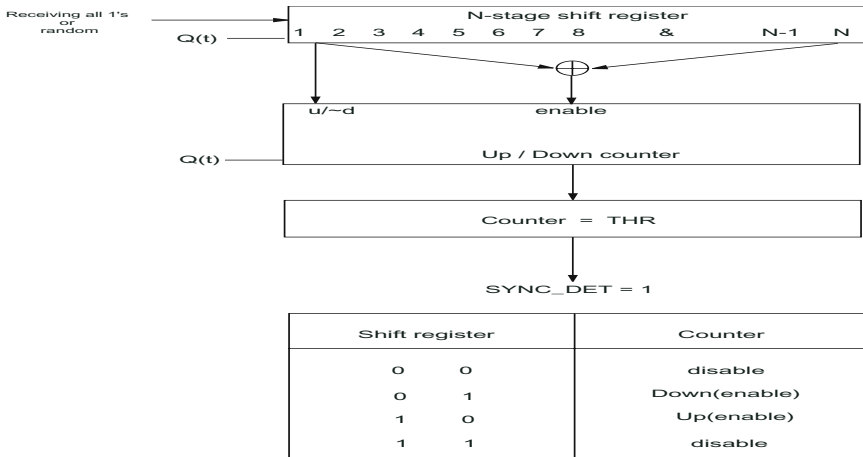


Fig. 6. Proposal of all bits "1" pattern detector

becomes high, if it is printed "1", the counter becomes decreased so it maintains the number of "1" that is included in register. In an output, when counter value compared with threshold (N_T), if the counter value is higher than threshold value we know that SYNC_DET=1 is printed.

2.2 The Encryption Algorithm

The way of encryption is divided into stream cipher, block cipher and Public key algorithm. There are four modes, ECB mode (electronic codebook), CFB mode (cipher feedback), CBC mode (cipher block chaining) and OFB mode (output feedback, in block cipher. In this chapter, we consider new design and selection problem in encryption algorithm which is already applied in MR. We have two approach ways to algorithm: one is OFB mode application way - it is changed Block cipher to alternative algorithm in order to meet the requirement wireless environment - another is to apply in Stream cipher.

2.2.1 Block Cipher Algorithms with OFB Mode

It is possible to apply OFB mode of Block cipher in wireless network with much noise. We could apply OFB mode to Block cipher by changing because the channel's quality become bad in case of applying to ECB mode, Most of the Block algorithm could apply to OFB mode. In this paper, we consider the algorithm which is to make up for the encryption weakness of DES algorithm which is already verified internationally. As the results of that, it is possible for FIPS-197 encryption AES algorithm [13], Triple-DES algorithm, ARIA algorithm and SEED algorithm[14] to be applied as show table 1.

To apply the wireless communication method, it needs to apply OFB mode. The applied OFB mode changes the encryption algorithm to PN-generator and

Table 1. Algorithm proposal of applied in enhanced encryption system

Previous algorithm		Improved algorithm	
Confidentiality algorithm	Identification algorithm	Confidentiality algorithm	Identification algorithm
DES	MD5	1) Block cipher: AES[13] ARIA, SEED[14], T-DES,IDEA 2) Stream cipher: LILI-II[15] Dragon, Parallel LM[16]	SHA-160 SHA-1, MD5, SEED-CBC AES-CBC T-DES-CBC, ARIA-CBC

uses it, then it returns output block to input registry, and generate PN-series. Finally it produces encryption by plain text and bit-by-bit XOR operation.

2.2.2 Stream Cipher Algorithms

LILI-II[15], Dragon, and Parallel LM[16] which are exemplified by Stream cipher algorithm are suitable for wireless communication. The first has a character of safe and high speed; the second is applied with the parallel Stream cipher method to achieve high speed for the existing LM generator.

2.3 Zero-Suppression Algorithm

In a synchronous stream cipher system, the proposed ZS algorithm [7], blocks (8) and (14), can suppress k or more zeros between successive ones in the ciphertext at the sender end and then completely recover the original message at the receiver end. This is useful in a system which limits the number of consecutive zeros, such as the T1-carrier system where k , the maximum number of permitted consecutive-zero bits, is 15.

3 System Performances

For a heavy noisy channel at $BER=0.1$, the window size $N=128$, threshold $N_T=25$ and $THR=128-25=103$ were designed/selected as the performance parameters, plus a probability of false-detection in sync. pattern $P_F=0.96667 \times 10^{-12}$, a probability of detection $P_D=0.9996387$, and a probability of missing $P_M=0.36 \times 10^{-3}$ were computed based on section II, as shown in tables 1 and 2. Although $BER=0.01$ is a very noisy channel, $P_D=1-10^{-15}$ was estimated as shown table 2. Therefore, the proposed system would appear to produce a highly reliable synchronization performance even in a noisy channel.

Therefore, the proposed system also possesses a highly secure keystream generator within a period of about 10^{42} , good randomness, an appropriate maximal linear complexity of about 10^{42} , and a maximal order correlation immunity of 4 with an M-L two-pass key agreement mechanism including mutual implicit

authentication. Its processing speed is fast approaching DS1 class(1.544 Mbps, 647ns/1-bit interval of the system clock) through DS3 class(44.736 Mbps, 22ns/1-bit output) for 50MHz system clock intervals. Finally, the proposed system can be implemented by either software or hardware and is appropriate for a link encryption model.

4 Conclusion

This paper proposed an improved initial synchronization method (referred to as "absolute synchronization") that can be applied to noisy channels (wireless) and produces a high-performance on the probability of synchronization. In addition, a stream cipher system was proposed for the absolute synchronization of keystream synchronization, a specified summation generator with 5-bit inputs plus 2-bit carries for data confidentiality, a zero-suppression algorithm for system stabilization. In summary, the proposed system includes a keystream generator that provides a high level of security over a period of about 10^{42} , good randomness, an appropriate maximal linear complexity of about 10^{42} with a two-pass key agreement mechanism including mutual implicit authentication. Its processing speed is also fast approaching DS1 class (1.544 Mbps, 647ns/1-bit interval) through DS3 class (44.736 Mbps, 22ns/1-bit interval) for 50MHz system clock intervals. Furthermore, even though $BER=10^{-2}$ is a very noisy channel, a highly reliable sync-detection probability of sharply one ($1-10^{-15}$) was achieved. Finally, the proposed system can be implemented using either software or hardware and is appropriate for a link encryption model.

References

1. B. Schneier, Applied Cryptography : Protocols, Algorithms, and Source Code in C, 2nd Ed., John Wiley and Sons, Inc., New York, USA, 1996.
2. A. Menezes, P. van Oorschot, S. Vanstone, Handbook of Applied Cryptography, CRC Press, 1997.
3. W. Diffie and M. E. Hellman, "New Directions in Cryptography," IEEE Trans. on Infor. Theory, Vol. IT-22, No. 6, pp. 644-654, Nov. 1976.
4. R. A. Rueppel, Analysis and Design of Stream Ciphers, Springer-Verlag, 1986.
5. Hoonjae Lee, SangJae Moon, "On An Improved Summation Generator with 2-Bit Memory," Signal Processing , Vol. 80, No.1, pp. 211-217, Jan. 2000.
6. M. Tatebayashi, N. Matsuzaki and D. B. Newman, "A Cryptosystem using Digital Signal Processors for Mobile Communication," ICASSP'90, pp. 37.1.1 - 37.1.4, 1990.
7. Hoonjae Lee, Sangjae Moon, "A Zero-Suppression Algorithm for the synchronous stream cipher," Applied Signal Processing, Vol.5, No.4, pp.240-243, 1998.
8. S. Moon, P. Lee, "A Propose of a Key Distribution Protocol," The Proceeding of The Korean Workshop on Information Security and Cryptography-WISC'90, pp. 117-124, 1990.
9. J. Proakis, Digital Communications (3rd Ed.), McGraw-Hill, Inc. 1995.
10. H.C.A. van Tilborg, Fundamentals of Cryptology, Kluwer Academic Publishers, 2000.

11. B. Park, H. Choi, T. Chang and K. Kang, "Period of Sequences of Primitive Polynomials," *Electronics Letters*, Vol. 29, No. 4, pp. 390-391, Feb. 1993.
12. H. J. Beker and F. C. Piper, *Secure Speech Communications*, Academic Press, London, 1985.
13. NIST, "*Announcing the Advanced Encryption Standard (AES)*", FIPS-197, Nov. 2001
14. KISA, *Development of SEED, the 128-bit block cipher standard*, Oct, 1998. (<http://www.kisa.or.kr>)
15. A. Clark, E. Dawson, J. Fuller, J. Golic, Hoon-Jae Lee, W. Millan, Sang-Jae Moon, L. Simpson, "*The LILI-II Keystream Generator*," LNCS 2384, pp.25-39, Jul. 2002 (ACISP'2002)
16. Hoonjae Lee, Sangjae Moon, "*Parallel Stream Cipher for Secure High-Speed Communications*," "*Signal Processing*", Vol.82,No.2,pp.259-265, Feb, 2002.

Recognition of Concrete Surface Cracks Using ART2-Based Radial Basis Function Neural Network

Kwang-Baek Kim¹, Hwang-Kyu Yang², and Sang-Ho Ahn³

¹ Department of Computer Engineering, Silla University, Busan 617-736, Korea
gbkim@silla.ac.kr

² Department of Multimedia Engineering, Dongseo University, Busan 617-716, Korea
hkyang88@hanmail.net

³ Department of Architectural Engineering, Silla University, Busan 617-736, Korea
shahn@silla.ac.kr

Abstract. In this paper, we proposed the image processing techniques for extracting the cracks in a concrete surface crack image and the ART2-based radial basis function neural network for recognizing the directions of the extracted cracks. The image processing techniques used are the closing operation of morphological techniques, the Sobel masking used to extract edges of the cracks, and the iterated binarization for acquiring the binarized image from the crack image. The cracks are extracted from the concrete surface image after applying two times of noise reduction to the binarized image. We proposed the method for automatically recognizing the directions (horizontal, vertical, -45 degree, 45 direction degree) of the cracks with the ART2-based RBF(Radial Basis Function) neural network. The proposed ART2-based RBF neural network applied ART2 to the learning between the input layer and the middle layer and the Delta learning method to the learning between the middle layer and the output layer. The experiments using real concrete crack images showed that the cracks in the concrete crack images were effectively extracted and the proposed ART2-based RBF neural network was effective in the recognition of the extracted cracks directions.

1 Introduction

Because the cracks in concrete structures have bad effects on the tolerance, the durability, the waterproof and the appearance of the structures, they bring about some worse problems in the structures. Therefore the causes of cracks must be accurately examined and the durability and the safety of the structures must be evaluated. If necessary, the repair and rehabilitation must be established. When we draw a deduction on the causes of cracks in the structures, the patterns and the distribution characteristics of cracks, these become important factors to judgment [1]. Because manual works by inspectors may be too subjective, techniques which enables objective examinations by computers is necessary [2], [3].

In this paper, we proposed the recognition method, which automatically extracts cracks from a surface image acquired by a digital camera and it also recognizes the directions (horizontal, vertical, -45 degree, and 45 degree) of the cracks using the

ART2-based RBF neural network. We compensate an effect of light on a concrete surface image by applying the closing operation, which is one of morphological techniques, extract edges of cracks by Sobel masking, and binarize the image by applying the iterated binarization technique [4]. Two separate times of noise reduction are applied to the binarized image for effective noise elimination. After minute noises are eliminated by using the average of adjacent pixels corresponding to a 3x3 mask, more noises are eliminated by analyzing the regular ratio of length and width with Glassfire labeling algorithm. The specific region of cracks is extracted from the noise-eliminated binarized image. We proposed the method to automatically recognize the directions of cracks by applying the ART2-based RBF neural network. The proposed ART2-based RBF neural network applied ART2 to the learning between the input layer and the hidden layer, and the Delta learning method is used in learning between the hidden layer and the output layer.

2 Crack Detection Using Image Processing Techniques

The overall process for the crack detection and recognition algorithm using the proposed techniques in this paper is described in Fig. 1. At first, cracks are extracted from a concrete surface image using some image processing techniques, and then, the directions of cracks are automatically recognized by applying the ART2-based RBF neural network, which is proposed in this paper.

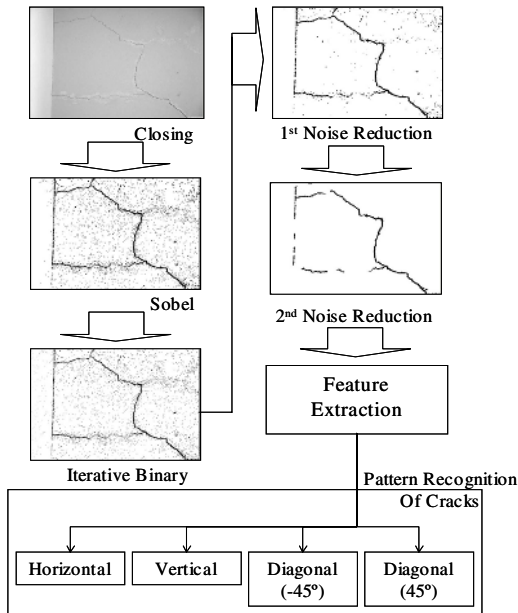


Fig. 1. Overview of proposed crack detection and recognition algorithm

2.1 Compensation for an Effect of Light

The brightness of the background varies according to the direction and the amount of light in an environment when photographing concrete surfaces. Sobel masking, which is sensitive to a value of brightness, cannot extract edges in the dark regions due to the effect of light. Therefore, for compensating effectively an effect of light, we applied the closing operation being one of the morphological techniques. The closing operation performs the dilation operation after the erosion operation. The dilation operation and the erosion operation are as follows:

$$(f \odot g)(x) = \max \{y : g(z-x) + y \ll f(z)\} \tag{1}$$

$$(f \oplus g)(x) = \min \{y : -g(-(z-x)) + y \gg f(z)\} \tag{2}$$

In Fig. 2, (a) is the original image and (b) is the closing image, which is generated by applying the closing operation to (a) and this shows appreciable cracks.

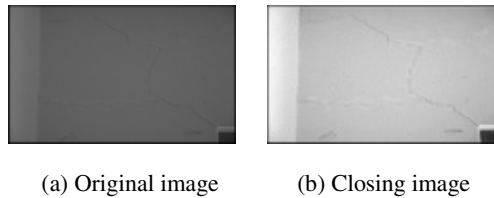


Fig. 2. Original image and closing image of a crack image

2.2 Crack Detection

Sobel masking is applied to the closing image for improving the performance of edge extraction based on features such as the great difference between the brightness of cracks and the brightness of the surface of a concrete structure. The edge extraction finds the change of brightness by the differential operator and the two masks shown in Fig. 3 is used for fast operation.

-1	0	1
-2	0	2
-1	0	1

Sobel-X

-1	-2	-1
0	0	0
1	2	1

Sobel-Y

Fig. 3. Sobel Mask

2.3 Binarization of a Crack Image

The iterated binarization selects the first estimated threshold value, it iterates an update of threshold value until the value doesn't change, and this is followed by it selecting the final threshold value.

Step 1. Select the first estimated threshold value T^0 .

Step 2. Divide the image into two regions R_1 and R_2 using the estimated threshold value T^t

Step 3. Calculate the average of the gray values, u_1 and u_2 , for each region.

$$u_1 = \frac{\sum f(i, j)}{N_1} \quad u_2 = \frac{\sum f(i, j)}{N_2}$$

N_1 and N_2 is the number of pixels in each region.

Step 4. Calculate the new threshold value.

$$T^{(t+1)} = \frac{u_1 + u_2}{2}$$

Step 5. Repeat Step 2 to 4 until the values of u_1 and u_2 cannot be changed.

Fig. 4. Iterated binarization algorithm

2.4 Noise Reduction

For eliminating noises without the influence of the cracks and the background, noise reduction operation is applied twice. Firstly, after the 3x3 mask is applied to the binarized image, if 1's pixels are more than 0's ones among the adjacent 9 ones, the center pixel is set to 1. Otherwise, the center pixel is set to 0 as shown in Fig. 5. This process eliminates minute noises.

0	0	0
0	0	1
0	1	0

1	1	1
1	1	1
1	0	0

Fig. 5. 3x3 Mask for noise reduction

Secondly, the Glassfire labeling technique is applied for eliminating additional noises. Glassfire labeling is the labeling method examining the adjacent pixels of the current pixel one by one recursively until all the adjacent pixels are labeled [5]. In the labeled image, the area of each labeled region is calculated by using the first pixel and last pixel of the region. In this paper, through the experiment, the criterion of the area

is set to 1.7. Therefore, if the ratio of the length and width is less than 1.7, they are determined as noises and are eliminated.

3 Crack Recognition Using the ART2-based RBF Neural Network

The learning of ART2-based RBF neural network is divided to two stages. In the first stage, competitive learning is applied as the learning structure between input layer and middle layer. And the supervised learning is accomplished between middle layer and output layer [7][8]. Output vector of the middle layer in the ART2-based RBF neural network is calculated by formula (3), and as shown in formula (4), the node having the minimum output vector becomes the winner node.

$$O_j = \frac{1}{N} \sum_{i=1}^N (|x_i - w_{ji}(t)|) \tag{3}$$

$$O_j^* = \text{Min}\{O_j\} \tag{4}$$

where $w_{ij}(t)$ is the connected weight value between input layer and middle layer.

In the ART2-based RBF neural network, the node having the minimum difference between input vector and output vector of the hidden layer is selected as the winner node of the middle layer, and the similarity test for the winner node selected is the same as formula (5).

$$O_j^* < \rho \tag{5}$$

where ρ is the vigilance parameter in the formula.

The input pattern is classified to the same pattern if the output vector is smaller than the vigilance parameter, and otherwise, to the different pattern. The connected weight is adjusted to reflect the homogeneous characteristics of input pattern on the weight when it is classified to the same pattern. The adjustment of the connected weight in ART2 algorithm is as follows:

$$w_{j^*i}^*(t+1) = \frac{w_{j^*i}^*(t) \cdot u_n + x_i}{u_n + 1} \tag{6}$$

where u_n indicates the number of updated patterns in the selected cluster.

The output vector of the middle layer is normalized by formula (7) and applied to the output layer as the input vector.

$$z_i = 1 - \frac{O_j}{N} \tag{7}$$

The output vector of the output layer is calculated by formula (8).

$$O_k = f \left(\sum_{j=1}^M w_{kj} \cdot z_j \right) \tag{8}$$

$$f(x) = \frac{1}{1 + e^{-x}} \tag{9}$$

The error value is calculated by comparing the output vector with the target vector. The connected weight is adjusted like formula (10) using the error value.

$$\delta_k = (T_k - O_k) \cdot O_k \cdot (1 - O_k) \tag{10}$$

$$w_{kj}(t + 1) = w_{kj}(t) + \alpha \cdot \delta_k \cdot z_j \tag{11}$$

4 Experiments and Performance Evaluation

The crack images are acquired by Sony’s Cyber-shot 5.0 digital camera. Fig. 6 is the result of the specific region extraction in a crack image and Fig. 7 is the result of the recognized directions of extracted cracks.

For analyzing the performance of the ART2-based RBF neural network, the extracted 25 crack patterns are used as input patterns. The vigilance parameter is set to 0.15, the number of output nodes to 6, and the learning rate to 0.5.

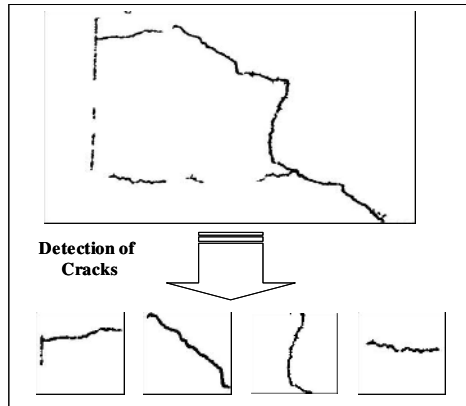


Fig. 6. Specific crack extraction in a crack image

Table 1 summarizes the performance measurement of the ART2-based RBF neural network. In Table 1, the failure cases in crack recognitions are the ones which uses enlarged or reduced images as input images, and the recognition of the non-directional cracks. Fig. 8 shows the graph for the change rate of TSS (Total Sum of Square) of error according to the number of epoch. As shown in Fig. 8, we know that the proposed algorithm has fast convergence and a good learning stability.

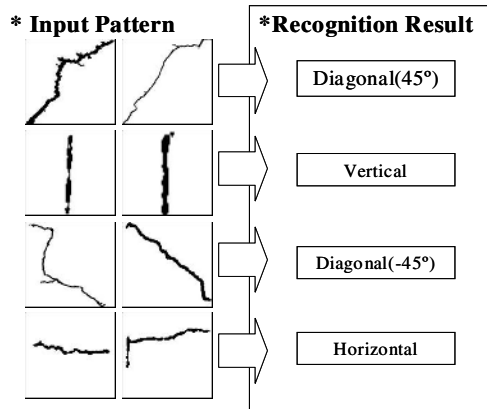


Fig. 7. Result of crack recognition

Table 1. Learning result of specific crack

	ART2-based RBF Neural Network
The number of node in the hidden layer	14
The number of Epoch	198
Recognition rate	24/25

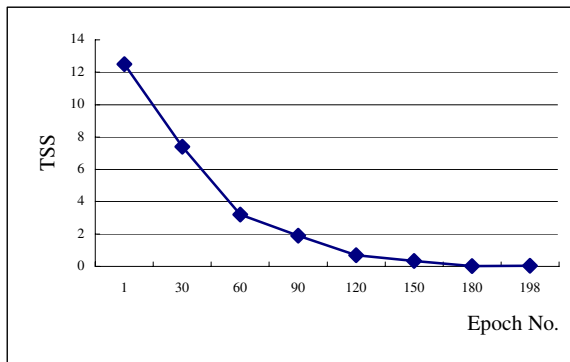


Fig. 8. Epoch vs. TSS

5 Conclusion

In this paper, we proposed the recognition method to automatically extract the cracks of a concrete surface image acquired by the digital camera and to recognize the

direction (horizontal, vertical, -45 degree, and 45 degree) of the specific cracks using the ART2-based RBF neural network. We compensate an effect of light on a concrete surface image by applying the closing operation, which is one of the morphological techniques, extract the edges of cracks by Sobel masking, and binarize the image by applying the iterated binarization technique. Noise reduction is applied twice to the binary image for effective noise elimination.

After the specific regions of cracks are automatically extracted from the preprocessed image by applying Glassfire labeling algorithm to the extracted crack image, the cracks of the specific region are enlarged or reduced to 30x30 pixels and then used as input patterns to the ART2-based RBF neural network. The learning of the ART2-based RBF neural network is divided to two stages. At the first stage the competitive learning is performed between input layer and middle layer, and at the second stage the supervised learning is performed between middle layer and output layer. The ART2-based RBF neural network shows the effectiveness of learning and recognition for the directions of the extracted cracks.

In this paper, when the enlarged or reduced images as inputs are used, the recognition of non-directional cracks fails. In future studies, this will be examined by the new algorithm which will recognize the non-directional cracks by extracting parameters from the features discovered through the patterns of the cracks.

References

1. Lee, B. Y., Kim, Y. Y. and Kim, J. K.: Development of Image Processing for Concret Surface Cracks by Employing Enhanced Binarization and Shape Analysis Technique. Journal of the Korea Concrete Institute, Vol. 17. No. 3. (2005) 361-368
2. Lee, B.Y., Park, Y. D. and Kim, J. K.: A Technique for Pattern Recognition of Concrete Surface Cracks. Journal of the Korea Concrete Institute, Vol. 17. No. 3. (2005) 369-374
3. Kim, Y. S. and Haas, C. T.: An Algorithm for Automatic Crack Detection, Mapping and Representation. KSCE Journal of Civil Engineering, Vol. 4. No. 2. (2000) 103-111
4. Gonzalez, R. C., Woods, R. E. and Eddins, S. L.: Digital Image Processing. Pearson Prentice Hall, (2004)
5. Pitas, I.: Digital Image Processing Algorithms and Applications. John Wiley & Sons INC, (2000)
6. Panchapakesan, C., Ralph, D. and Palaniswami, M.: Effects of Moving the Centers in an RBF Network. Proceedings of IJCNN, Vol. 2. (1998) 1256-1260
7. Kim, K. B., Joo, Y. H. and Cho, J. H.: An Enhanced Fuzzy Neural Network. Lecture Notes in Computer Science, LNCS 3320. (2004) 176-179
8. Pandya A. S., and Macy R. B.: Neural Networks for Pattern Recognition using C++. IEEE Press and CRC Press, (1995)

Hybrid Image Mosaic Construction Using the Hierarchical Method*

Oh-Hyung Kang, Ji-Hyun Lee, and Yang-Won Rhee

Department of Computer Science, Kunsan National University,
68, Miryong-dong, Kunsan, Chonbuk 573-701, South Korea
{ohkang, jhlee, ywrhee}@kunsan.ac.kr

Abstract. This paper proposes a tree-based hierarchical image mosaicing system using camera and object parameters for efficient video database construction. Gray level histogram difference and average intensity difference are proposed for scene change detection of input video. Camera parameter measured by utilizing least sum of square difference and affine model, and difference image is used for similarity measure of two input images. Also, dynamic objects are searched through macro block setting and extracted by using region splitting and 4-split detection methods. Dynamic trajectory evaluation function is used for expression of dynamic objects, and blurring is performed for construction of soft and slow mosaic image.

1 Introduction

Mosaic image construction refers to the creation of single new image that compose several videos and still images related together [1]. Mosaic is classed by static mosaic that focus on background, dynamic mosaic for dynamic object description, and synopsis mosaic that appear representatively integrating static and dynamic description [2].

Technologies of mosaic construction process consist of arrangement, integration and overlapping analysis of continuous images. The most traditional field of application for panoramic mosaic system is construction of air artificial satellite picture, recently scene fixing and scene sensing, video compression and indexing, and research of camera resolution as well as, even study in very various field to simple picture edit. Frame image that represent one scene in video is known as representative frame image, it have many difficulty in understanding scene of moving picture only by representative frame. Mosaic image solves ambiguity problem of information that can drop in representative frame of video because it make resemble much images to single image. Mosaic image has great advantage that is use of minimum storage space, fast transmission of data, and understanding of whole scene because it is including all adjacent images.

Many papers suggested difference image, coordinate conversion technique of image, optical flow and problem and solution of motion estimation. Philip et al. [3]

* This research was supported by the Program for the Training of Graduate Students in Regional Innovation which was conducted by the Ministry of Commerce, Industry and Energy of the Korean Government.

proposes a description method of mosaic image that each input image is concerned with transformation matrix that is not method to reflect on general plane. However, this method has a disadvantage that is very sensitive to noise. Xiong et al. [4] proposed way to construct the virtual world by getting 4 images rotating camera 90 degrees. IBM's system called ImageMiner is a system that integrate still image analysis in video analysis as IBM's trademark, described method to recognize object using color, texture and shape information as step that analyze mosaic image [5]. However, ImageMiner system has problem that it does not recognize dynamic object because mosaic image is recognized object only with color, texture, and shape information. JuHyun et al. [6] proposed about dynamic mosaic construction for dynamic object, but specific method that can identify camera and dynamic object that move did not describe.

The easiest method to create mosaic image that is consisted of transfer only between images. Images moved, so it can be implemented easily on minimum restriction, and mosaic of high resolution that quality is high of mosaic image can be created, and the advantage is that the computing time is fast [7].

Camera motion must be measured necessarily to construct mosaic. One of the camera motion estimation is a method that performs through parameter calculation of camera that uses optical flow [8]. Also, motion models that used to extract camera parameter are used as two dimensions parameter motion model and complicated three-dimension motion model [9]. Affine model that can measure rotation at the same time including movement and scaling among 2 dimensions motion model is utilized most in camera parameter measure.

This paper proposes, after preprocessing of input as still image and video frame, method that measure parameters of camera and dynamic objects, and method that construct mosaic image by object extraction. Whole system structure is shown in Figure 1.

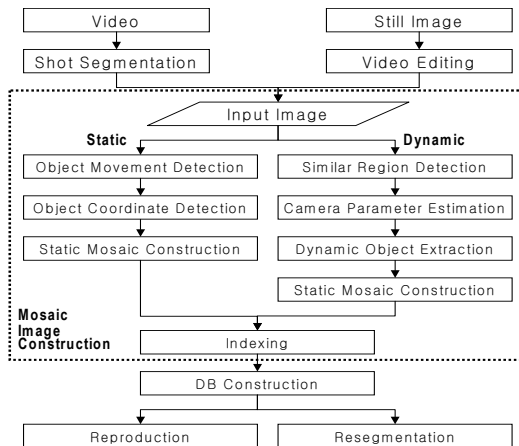


Fig. 1. Overall system structure

Chapter 2 presents about the scene change detection of input video. Chapter 3 explains about hierarchical mosaic image construction method and Chapter 4 explains about static and dynamic mosaic image construction. Chapter 5 experiments static and a dynamic mosaic image construction and Chapter 6 concludes this paper and cite about the future work.

2 Scene Change Detection of Video

In this paper, scene change detection is based on the framework of the twin-threshold method that is able to detect both abrupt and gradual transition. To segment the video, suitable metrics is defined, so that a shot boundary is declared whenever that metric exceeds a given threshold. The Color- χ^2 intensity histogram difference is used as the first metric in our algorithm because histogram is more powerful to object motion than other metrics, it is shown in Equation 1.

$$D(I_t, I_{t-1}) = \frac{1}{3} \bullet \sum_{j=1}^N \left(\frac{(H_t^r(j) - H_{t-1}^r(j))^2}{H_t^r(j)} \times 0.333 + \frac{(H_t^g(j) - H_{t-1}^g(j))^2}{H_t^g(j)} \times 0.333 + \frac{(H_t^b(j) - H_{t-1}^b(j))^2}{H_t^b(j)} \times 0.333 \right) \quad (1)$$

The second metric is the average intensity difference, it is shown in Equation 2.

$$AI_i = \frac{\sum_{j=1}^{Bins} j * H_i(j)}{\sum_{j=1}^{Bins} H_i(j)} \quad AI_{i-1} = \frac{\sum_{j=1}^{Bins} j * H_{i-1}(j)}{\sum_{j=1}^{Bins} H_{i-1}(j)} \quad AD_i = AI_i - AI_{i-1} \quad (2)$$

$H_t^r(j)$, $H_t^g(j)$, and $H_t^b(j)$ are the bin values of the histogram of t 'th frame in red, green and blue color channels. $D(I_t, I_{t-1})$ denotes the intensity histogram difference between frame t and its preceding frame ($t-1$). AI_i is the average color value of the frame i and AI_{i-1} is the average color value of the frame ($i-1$). AD_i is the average color difference values between frames i and ($i-1$).

3 Hierarchical Image Mosaicing

Requirement that is gone ahead first in mosaic image creation is to do all images to arrange resemble images consecutively. This needs to match overlapping area between two images to compare to one part.

3.1 Mosaicing Flow

In mosaic system as shown in Figure 2, image in scene that acquired by video camera and arranged image that acquired by general camera are used for input. Image that is

acquired by general camera can have big difference between each image. In this case, it is effective to construct mosaic through similarity measure.

Figure 2 shows the flow for mosaic image construction. After videos of real life that is acquired from camera is input by continuous frame as segmented scene and still images are input automatically arranged image, then two frames or images are compared. For measure camera movement between two images, least sum of square difference and affine model is used. In this paper, macro block setting, region splitting, and 4-split detection method are proposed to extract and recognize dynamic object.

After extraction and recognition of the dynamic object, construct smooth mosaic image generally applying trajectory description of dynamic object and blurring method.

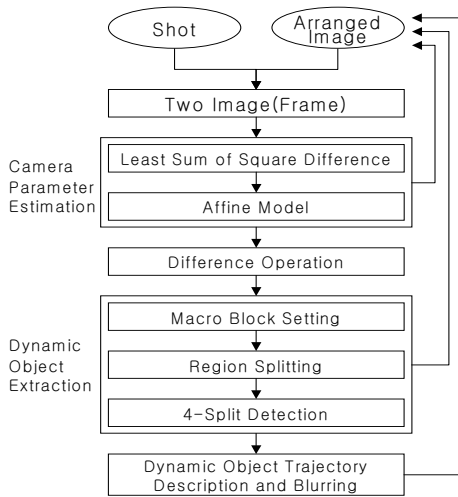


Fig. 2. Mosaicing Flow

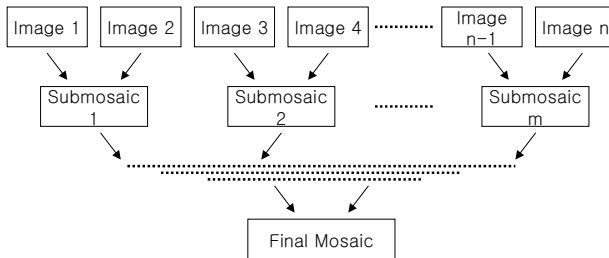


Fig. 3. Tree-based Mosaic Image

3.2 Tree-Based Mosaicing

This paper proposes creation method of tree-based mosaic image to get fast computing time of mosaic image creation. Tree-based mosaic image creation is not that adja-

cent images are created to consecutively mosaic image. Image mosaicing creates one partial mosaic image by comparing only two of adjacent images. Therefore all source images are created to partial mosaic image. One general mosaic image is created by the creation of new partial mosaic image repeatedly by such created partial mosaic image. If source image is 16, then 4 levels of partial mosaic exist.

4 Static and Dynamic Image Mosaic

4.1 Static Image Mosaicing

The method to extract dynamic object in static background uses difference image technique, it is calculated by using difference operation between pixels of two images as shown in Equation 3.

$$D(x, y) = |I_a(x, y) - I_b(x, y)| \quad (3)$$

Here I_a is a base image and I_b is a reference image. Also, it needs local detection to detect object region. In this paper, it detects object using 16×16 macro block, detect by performing local difference operation as shown in Equation 4.

$$LD(x, y) = \sum_{y=1}^N \sum_{x=1}^N |I_a(x, y) - I_b(x, y)| \quad (4)$$

Local difference operation is to subtract reference image I_b from base image I_a and calculates this as size of 16×16 pixel. In this way, after detecting object region by performing local difference operation using macro block, similarity of two images is measured by using least sum of square difference, and measure affine parameter using coordinate value three point or more obtained by calculating camera movement, then two images are matched for constructing static mosaic image. Least sum of square difference and affine parameter measure is described in detail in mosaic construction using dynamic background and object in the next section.

4.2 Dynamic Image Mosaicing

There is a difficulty to detect movement of camera and object between two images that have dynamic object. Even if there is a camera movement, if object is filling image as a whole, it do not recognize camera motion. Also, if big object moves when camera does not move, it can't recognize this as camera motion.

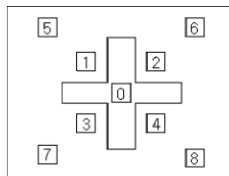


Fig. 4. Macro Block of Base Image

This paper proposes a method to solve this ambiguity. Simple method that can think most usually is that object is occupying middle of screen mainly when camera takes an important object. So, similarity is measured based on macro block of outer region except middle of base image in this method.

Figure 4 shows macro block of base image for comparison. Each macro block selects macro block by non-linear as central outer region of cross shape. Preferential assumption is that the size of object does not occupy half of image. The algorithm that detects dynamic object by setting macro block is shown below.

- (1) Input two image (frame);
- (2) Similarity measure of macro block 1, 2, 3 and 4;
If (value of least sum of square difference < threshold)
Then {measure of motion vector; goto (3);}
Else goto (1);
- (3) If (motion vector is similar to all four macro block)
Then {camera parameter = motion vector; goto (4);}
Else {measure of motion vector by extending to
macro block 5, 6, 7 and 8; goto (4);}
- (4) Measure of camera parameter through affine transformation;
- (5) Local difference operation between two images

If the results that get performing difference operation using above algorithm is more than threshold, it is considered that dynamic object exists. Next is to look around each method that is used in above algorithm.

4.2.1 Least Sum of Square Difference

First, it must extract correct camera parameter to look for similarity between two images. This paper proposes least sum of square difference about fixed window block as shown in Equation 5.

$$E(C) = \sum_{b \in W} [I_i(X + b) - I_j(X + b + d_k)]^2 \quad (5)$$

In equation 5, X refers to pixel position of x and y , and b refers to windows of a regular square of an image. Value of least sum of square difference is calculated through block of all d_k of reference image I_j in place that is calculated by square of difference value between reference image I_j and base image I_i . At this time, minimum value among square difference that is calculated is selected.

4.2.2 Affine Model

Camera parameter including rotation, scaling and movement of image is measured by using affine model based on the most similar pixel value that is detected by using Equation 6.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} \quad (6)$$

4.2.3 Local Difference Operation

Next, it must detect movement of object, local detection must perform for this. Method to present in this paper performs by comparison through difference operation

using motion vector and macro block between two images with camera motion that is calculated, as show in Equation 7.

$$E(O) = \sum_{b \in W} [I_i(X + b) - I_j(X - u(k) + b + d_k)]^2 \tag{7}$$

Here $u(k)$ refers to distance of camera movement as motion vectors of x axis and y axis, and error value between two images is calculated by subtraction of it. At this time, large movement is detected if the threshold is big and it detects the movement of object, while small motion can be extracted if the threshold is small.

4.2.4 Region Splitting

Also, this paper proposes a region splitting method through basic assumption for dynamic object detection. In Figure 5, if suppose that b and c region of two images show similar region between two images, following assumption is followed for detecting dynamic object. First, if some part of a region and some part of region c or d is corresponded, this region becomes object. Second, if some part of region b and d is corresponded, this region becomes dynamic object. Finally, conflicting region between region b and c that corresponding between two images becomes object.



Fig. 5. Region Splitting for Dynamic Object Detection

4.2.5 4-Split Detection

Dynamic object extraction is to extract only dynamic object by detection of similar region existing in dynamic object in two images. If the value of difference image between two similar region is large, there is an assumption that dynamic object existed within two images. As shown in Figure 6, region of 1st quadrant is calculated and compared. If value of difference image is small, computation proceeds to 2nd quadrant and so on continuously. Size of dynamic object ignores object fewer than smallest 7×7 pixel size, and when multiple dynamic objects are detected, the largest dynamic object between them is extracted. Detection process is performed up to last 8×8 block and dynamic object region is created by sum of blocks.



Fig. 6. 4-Split Detection Method

4.2.6 Dynamic Trajectory Description and Blurring

Background image composition compose only remainder background image after dynamic object extraction. At this time, background part of other remainder image is inserted on part of extracted dynamic object. After creating background mosaic image, the description of dynamic object express object that distance is more than 1.5 times of maximum width and height size of the extracted object. Evaluation function that is presented in this paper is shown in Equation 8.

$$\begin{aligned}
 &\text{if } (A > 1.5B) \text{ then describe dynamic object} \\
 &\text{where } A=Ii(x2, y2)-Ii-1(x1, y1) \\
 &\quad B=\text{Length}(O_{MAX}[(x_1, x_2), (y_1, y_2)])
 \end{aligned} \tag{8}$$

In Equation 8, after calculating maximum size (O_{MAX}) of dynamic object, when dynamic object moved to right, it is described if distance difference ($(Ii(x2, y2) - Ii-1(x1, y1))$) between left region of present image and right region of preceding image is more than 1.5 times of maximum object size.

Blurring creates visually smoothing mosaic image using the most general method that sum of whole mask set to 1 using 3 x 3 masks on the border part where each image may be integrated.

5 Experimental Result

To implement mosaic system that is proposed in this paper, image is acquired from digital zoom camera. Input image is used after normalizing of 320 * 240 sizes and implementation is performed by using Visual C++6.0 in Pentium-4 3.0GHz. Figure 7 display some of whole frame in 30 seconds video that takes interior of laboratory.

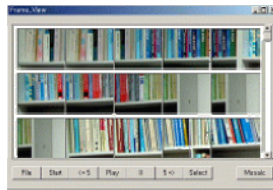


Fig. 7. Frame Images

Figure 8 displays 6 images that are selected arbitrarily among whole frame of Figure 7. These are images that are used in input to create mosaic image.



Fig. 8. Selected Input Image

Figure 9 is static mosaic image that is created from 6 input images which appeared in Figure 8. In Figure 9, we can know that image length from left to right is prolonged fairly as it is panoramic mosaic image that dynamic object does not exist. Through this mosaic image, we can understand easily contents of whole video.



Fig. 9. Static Mosaic Image

Experiment that construct dynamic mosaic image in video that dynamic object exists is as following. First, Figure 10 is a window that extract key frame by proposed scene change detection method for a video 30 seconds long that dynamic object exists in it.

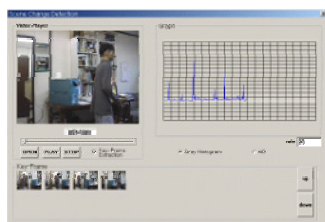


Fig. 10. Scene Change Detection

Figure 11 is dynamic mosaic image describing dynamic object. When dynamic object is moved more than 80 pixels in an experiment, dynamic object appeared in mosaic image.

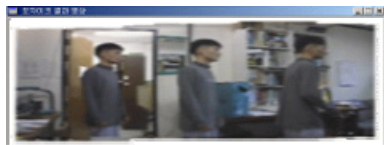


Fig. 11. Dynamic Mosaic Image

Result that construct mosaic image for 5 videos 30 seconds long that captured in campus is as Table 1.

Table 1. Mosaic Image Construction Result

Input Video		Mosaic Image		Reduction Ratio
Average Frame Number	Average Volume	Average Frame Number	Average Volume	
900	8MB	97	750KB	90.63%

Constructed mosaic image reduces storage space of 90.63% than input video's volume. That is, it does so that can understand whole video's contents by one image of very small volume.

In this paper, unique and notable characteristics of research and experiment that compare with existing mosaic system are summarized in Table 2.

Table 2. System Compare and Evaluation

Item	Existing system	Proposed system
Construction base	Consecutive frame-based	Tree-based
Utilizing type	Static or dynamic	Both static and dynamic
Computation area	Whole image	Partial image
Parameter	Mainly camera parameter	Both camera and object parameter
Trajectory description	Ambiguous	Clear and flexible
Mosaic image border	No-soft	Soft

6 Conclusion

This paper proposed way to construct tree-based hierarchical image mosaicing using camera and object parameter for efficient video database construction. Camera parameter measured by utilizing least sum of square difference and affine model, and difference image is used for similarity measure of two input images. Also, dynamic objects are detected through macro block setting and extracted by using region splitting and 4-split detection methods. And soft mosaic image is constructed through blurring after used dynamic trajectory evaluation function for expression of dynamic object. This tree-based mosaic system presented foundation that can reduce time and storage space to construct video database.

In the future, consecutive research about efficient scene change detection, object detection and image reappearance is needed for creating enhanced mosaic image.

References

1. Richard Szelisk, Shum H.: Creating Full View Panoramic Image Mosaics and Environment Maps, In Proc. of SIGGRAPH, (1997) 251-258.
2. Shaolei Feng, Hanqing Lu and Songde Ma : Mosaic representations of video sequences based on slice image analysis, Pattern Recognition Letters, Vol. 23, No. 5, (2002) 513-521
3. Philip F. McLauchlan and Allan Jaenicke : Image mosaicing using sequential bundle adjustment , Image and Vision Computing, Vol. 20, No. 9, (2002) 751-759

4. Xiong Y., Turkowski K.: Creating Image-Based VR Using A Selfcalibrating Fisheye Lens, Proc. CVPR' 97, (1997) 237-243.
5. Krey B J., Roper M., Alshuth P., Hermes Th., Herzog O.: Video Retrieval by Still-Image Analysis with ImageMiner, SPIE, (1997) 36-44.
6. JuHyun Cho and SeongDae Kim : Object detection using multi-resolution mosaic in image sequences , Signal Processing: Image Communication, Vol. 20, No. 3, (2005) 233-253
7. Aya Aner-Wolf and John R. Kender : Video summaries and cross-referencing through mosaic-based representation, Computer Vision and Image Understanding, Vol. 95, No. 2, (2004) 201-237
8. Udhav Bhosle, Sumantra Dutta Roy and Subhasis Chaudhuri : Multispectral panoramic mosaicing , Pattern Recognition Letters, Vol. 26, No. 4, (2005) 471-482
9. Jun-Wei Hsieh : Fast stitching algorithm for moving object detection and mosaic construction, Image and Vision Computing, Vol. 22, No. 4, (2004) 291-306

Public Key Encryption with Keyword Search Based on K-Resilient IBE

Dalia Khader

University of Bath, Department of Computer Science
ddk20@bath.ac.uk

Abstract. An encrypted email is sent from Bob to Alice. A gateway wants to check whether a certain keyword exists in an email or not for some reason (e.g. routing). Nevertheless Alice does not want the email to be decrypted by anyone except her including the gateway itself. This is a scenario where public key encryption with keyword search (PEKS) is needed. In this paper we construct a new scheme (KR-PEKS) the K-Resilient Public Key Encryption with Keyword Search. The new scheme is secure under a chosen keyword attack without the random oracle. The ability of constructing a Public Key Encryption with Keyword Search from an Identity Based Encryption was used in the construction of the KR-PEKS. The security of the new scheme was proved by showing that the used IBE has a notion of key privacy. The scheme was then modified in two different ways in order to fulfill each of the following; the first modification was done to enable multiple keyword search and the other was done to remove the need of secure channels.

1 Introduction

Bob wants to send Alice confidential emails and in order to ensure that no one except her can read it, he encrypts the emails before sending them so that Alice, and her alone, will possess the capability of decrypting it. Consider the scenario where Alice would like to download only the urgent emails to her mobile, leaving the rest to check later from her computer. Alice requires access to the email server (gateway) so that she can search her emails for the keyword “urgent” prior to downloading them. The server should facilitate the search for this keyword without being able to decrypt Alice’s private emails.

This scenario was first introduced in Boneh et al.’s paper [4]. They presented a general scheme called PEKS where Alice gives trapdoors for the words she wants the gateway to search for. The trapdoors come in the form of some kind of data that is used to test the existence of keywords within an email without revealing any other information(Section 3).

In [4] the authors constructed several schemes based on different security models but these schemes either had some limitation on the number of keywords to search for or were not secure enough (ie. were proven secure in random oracle).

In [2] the authors pointed out two important features that were not covered in [4]. The first one was the ability to search for multiple keywords. The second

characteristic, put forwards in Section 3.1 of this paper, was elimination of the requirement of secure channels, for sending trapdoors. These two new features issues are explained in details in the paper and new schemes to facilitate them will be introduced in this report paper(Section 3.6 and 3.7).

The ability of constructing IBE from PEKS was explored in [4]. IBE is a public key encryption where the public key is a direct product of the identity of the user [12] [5]. Building a PEKS from an IBE needs the latter to have some extra properties such as the notion of key privacy. Key privacy is a security property that implies that an adversary should not be able to guess which ID from a set of ID's was used in encrypting some email(Section 3.3).

Heng and Kurosawain [10] proposed a scheme called K-resilient IBE, this scheme does not lead to breach of privacy [1]. The basic K-resilient IBE was used to construct a PEKS that is fully secure (Section 3.4).

The next Section entails some important concepts that form the foundation for our work. Following that, Section 3 explains the PEKS scheme, its security notions, the relation with IBE, and last but not least the construction of a new PEKS scheme using the K-Resilient IBE was introduced. The last Section in this paper concludes the results of our work.

2 Preliminaries

In this Section we will go through some definitions that will be used further in this document.

2.1 Decisional Diffie-Hellman

The Decisional Diffie Hellman (DDH) Problem [7] is the ability to distinguish between $\langle g, g^a, g^b, g^{ab} \rangle$ and $\langle g, g^a, g^b, T \rangle$ where $a, b, c \in \mathbb{Z}_q, g \in G_q, G_q$ is a group of prime order q , and T is a random element that belongs to G_q .

The quadruple $\langle g, g^a, g^b, g^{ab} \rangle$ is called the real quadruple and the quadruple $\langle g, g^a, g^b, T \rangle$ is called the random quadruple. So if we have an adversary D that takes $X, Y, T \in G_q$ and returns a bit $d \in \{0, 1\}$, consider the following two experiments.

For both experiments $X \leftarrow g^x; Y \leftarrow g^y$

• $Exp_{G_q, D}^{ddh-real}$

$T \leftarrow g^{xy}$

$d \leftarrow D(q, g, X, Y, T)$

Return d

• $Exp_{G_q, D}^{ddh-rand}$

A random $T \in G_q$

$d \leftarrow D(q, g, X, Y, T)$

Return d

The advantage of the adversary D in solving the Diffie-Hellman is defined as follows:-

$$Adv_{G_q, D}^{ddh} = Pr[Exp_{G_q, D}^{ddh-real} = 1] - Pr[Exp_{G_q, D}^{ddh-rand} = 1]$$

If this advantage is negligible for any adversary D then we say DDH is hard to solve. The DDH was used in PEKS Section 3.4.

2.2 Hash Functions

A family of hash functions $\mathcal{H} = (G; H)$ is defined by two algorithms [9]. G is a probabilistic generator algorithm that takes the security parameter k as input and returns a key K . H is a deterministic evaluation algorithm that takes the key K and a string $M \in \{0, 1\}^*$ and returns a string $H_k(M) \in \{0, 1\}^{k-1}$.

Definition. Let $\mathcal{H} = (G; H)$ be a family of hash functions and let C be an adversary [9]. We consider the following experiment:

Experiment $Exp_{\mathcal{H}, C}^{cr}(k)$

$K \leftarrow G(k); (x_0; x_1) \leftarrow C(K)$

if $((x_0 \neq x_1) \wedge (H_K(x_0) = H_K(x_1)))$ then return 1 else return 0.

We define the advantage of C via $Adv_{\mathcal{H}, C}^{cr}(k) = Pr[Exp_{\mathcal{H}, C}^{cr}(k) = 1]$:

The family of hash functions \mathcal{H} is collision-resistant if the advantage of C is negligible for every algorithm C whose time-complexity is polynomial in k .

3 Public Key Encryption with Keyword Search

An encrypted email is sent from Bob to Alice [4]. The gateway wants to check whether a certain keyword exists in an email or not for some reason (for example routing). Nevertheless Alice does not want the email to be decrypted by anyone except her, not even the gateway itself. This is a scenario where public key encryption with keyword search (PEKS) is needed. PEKS encrypts the keywords in a different manner than the rest of the email. The gateway is given “trapdoors” corresponding to particular keywords. Using the PEKS of a word and trapdoor of a keyword, the gateway can test whether the encrypted word is the particular keyword or not.

General Scheme. According to [4] a PEKS consists of four algorithms as described below :

- **KeyGen**(s): Take a security parameter s and generate two keys a public key A_{pub} and private key A_{priv}
- **PEKS**(A_{pub}, W): It produces a searchable encryption for a keyword W using a public key A_{pub}
- **Trapdoor**(A_{priv}, W): Produce a trapdoor for a certain word using the private key.
- **Test**(A_{pub}, S, T_w): Given the public key A_{pub} , some searchable encryption S where $S = PEKS(A_{pub}, W')$, and the trapdoor T_w to a keyword W . Determine whether or not the word we are looking for W and the word encrypted W' are equal.

So Bob sends Alice through the gateway the following:

$$[E(A_{pub}, M), PEKS(A_{pub}, W_1), PEKS(A_{pub}, W_2), \dots, PEKS(A_{pub}, W_m)]$$

where $PEKS(A_{pub}, W_i)$ is a searchable encryption of the keywords and $E(A_{pub}, M)$ is a standard public key encryption of the rest of the message M .

3.1 Security Notions Related to PEKS

Security Under a Chosen Keyword Attack (CKA). For a PEKS to be considered secure we need to guarantee that no information about a keyword is revealed unless the trapdoor of that word is available [4]. To define security against an active adversary A we use the following game between A and challenger.

- **CKA-Setup:** The challenger runs the key generation algorithm and gives the A_{pub} to adversary A and keeps A_{priv} to itself.
- **CKA-Phase 1:** A asks the challenger for trapdoors corresponding to keywords of its choice.
- **CKA-Challenge:** The adversary decides when phase 1 ends. Then it chooses two words W_0, W_1 to be challenged on. The two words should not be among those for which A obtained a trapdoor in phase 1. The challenger picks a random bit $b \in \{0, 1\}$ and gives attacker. $C = PEKS(A_{pub}, W_b)$.
- **CKA-Phase 2:** A asks for more trapdoors like in phase 1 for any word of its choice except for the W_0, W_1 .
- **CKA-Guess:** A outputs its guess of b' and if $b' = b$ that means A guessed the encrypted message and the adversary wins.

We say that the scheme is secure against a chosen keyword attack (CKA) if A has a low advantage of guessing the right word being encrypted.

Secure Channels. In the PEKS scheme in [4] there is a need to have a secure channel between Alice and the server, so that an eavesdropper (Eve) can not get hold of the trapdoors sent. No one but the server should be capable of testing emails for certain keywords. This is one of the drawback that the authors of [2] tried to solve by generating a public and a private key that belong to the server. The PEKS algorithm was modified to encrypt keywords using both Alice's and the server's public key, while the testing algorithm needs the server's private key as an input. In this way the scheme is secure channel free (SCF-PEKS) because Eve can not obtain the server's private key, therefore can not test.

The SCF-PEKS is said to be IND-SCF-CKA secure when it ensures that the server that has obtained the trapdoors for given keywords cannot tell a PEKS ciphertext is the result of encrypting which keyword, and an outsider adversary that did not get the server's private key cannot distinguish the PEKS ciphertexts, even if it gets all the trapdoors for the keywords that it queries.

3.2 Handling Multiple Keywords

Multiple Keyword search in a PEKS is the capability of searching for more than one word either disjunctively or conjunctively. In PEKS [4] the only way to do this is to search for each word separately and then do the disjunctive or conjunctive operations on the result of the testing algorithm. This technique is impractical when it comes to a large number of keywords in one conjunctive search request, because every email is searched for every single keyword. In [8] a new scheme was suggested for conjunctive search called PECK. The scheme substitutes the PEKS algorithm with a PECK algorithm that encrypts a query of keywords. The testing is done with a trapdoor for each query instead of each word. So Bob sends Alice the following:

$$[E(A_{pub}, M), PECK(A_{pub}, (W_1, W_2, \dots, W_m))]$$

We say that the scheme is secure against a chosen keyword attack (CKA) if an adversary has a low advantage in guessing the right query of keywords being encrypted.

3.3 The Strong Relation Between IBE and PEKS

In [4] the authors showed how the algorithms used in IBE can be used for constructing a PEKS. They showed how with the four algorithms in an IBE Setup, Extract, Encrypt, and Decrypt [12] [5] could be used for to achieve the purpose of a PEKS scheme. So if keyword was used in place of an ID in the IBE scheme. The Setup algorithm will be equivalent to the KeyGen algorithm in a PEKS. Extracting the private key in the IBE will be in replace of generating trapdoors for keywords in PEKS scheme. Now if the Encryption algorithm in the IBE was used to encrypt some zero string of a certain length, the result will be a PEKS ciphertext that could later be tested by using the Decryption algorithm in an IBE and checking whether the result is the same string of zeros. The problem is that the ciphertext could expose the public key (W) used to create it. So we need to derive a notion of key privacy [3] for IBE to ensure that the PEKS is secure under a chosen keyword attack. Key privacy is a security notion first introduced in [3]. If an adversary can not guess which ID of a set of ID's in an IBE scheme was used in encrypting a particular ciphertext then that IBE scheme does not lead to breach the privacy of its keys. This could be under chosen plaintext attack(IK-CPA) or chosen ciphertext attack(IK-CCA).

3.4 Construction of a PEKS from the K-Resilient IBE (KRPEKS)

Since the K-resilient IBE scheme suggested in [10] is said to be IK-CCA secure ([1] for proof) it was tempting to construct a PEKS using that scheme. As any other PEKS scheme there are four algorithms, summarized in the following.

– KRPEKS-KeyGen

Step 1: Choose a group G of order q and two generators g_1, g_2

Step 2: Choose 6 random k degree polynomials where the polynomials are chosen over Z_q

$$P_1(x) = d_0 + d_1x + d_2x^2 + \dots + d_kx^k; P_2(x) = d'_0 + d'_1x + d'_2x^2 + \dots + d'_kx^k$$

$$F_1(x) = a_0 + a_1x + a_2x^2 + \dots + a_kx^k; F_2(x) = a'_0 + a'_1x + a'_2x^2 + \dots + a'_kx^k$$

$$h_1(x) = b_0 + b_1x + b_2x^2 + \dots + b_kx^k; h_2(x) = b'_0 + b'_1x + b'_2x^2 + \dots + b'_kx^k$$

Step 3: For $0 \leq t \leq k$; Compute $A_t = g_1^{a_t} g_2^{a_t}, B_t = g_1^{b_t} g_2^{b_t}, D_t = g_1^{d_t} g_2^{d_t}$

Step 4: Choose a random collision resistant hash function H (Section 2.2)

Step 5: Choose a random collision resistant hash function H' (Section 2.2)

Step 6: Assign $A_{priv} = \langle F_1, F_2, h_1, h_2, P_1, P_2 \rangle$

$$A_{pub} = \langle g_1, g_2, A_0, \dots, A_k, B_0, \dots, B_k, D_0, \dots, D_k, H, H' \rangle$$

– KRPEKS

Step 1: Choose a random $r_1 \in Z_q$

Step 2: Compute $u_1 = g_1^{r_1}; u_2 = g_2^{r_1}$

Step 3: Calculate for each keyword w

$$A_w \leftarrow \prod_{t=0}^k A_t^{w^t}; B_w \leftarrow \prod_{t=0}^k B_t^{w^t}; D_w \leftarrow \prod_{t=0}^k D_t^{w^t}$$

Step 4: $s \leftarrow D_w^{r_1}$

Step 5: Using the -exclusive or- operation calculate $e \leftarrow (0^k) \otimes H'(s)$

Step 6: $\alpha \leftarrow H(u_1, u_2, e)$

Step 7: $v_w \leftarrow (A_w)^{r_1} \cdot (B_w)^{r_1 \alpha}$

Step 8: $C \leftarrow \langle u_1, u_2, e, v_w \rangle$

– KRPEKS-Trapdoor

Run Extract of the IBE and the output is the trapdoor

$$T_w = \langle F_1(w), F_2(w), h_1(w), h_2(w), P_1(w), P_2(w) \rangle.$$

– KRPEKS-Test

Step 1: $\alpha \leftarrow H(u_1, u_2, e)$

Step 2: Test if $v_w \neq (u_1)^{F_1(w)+h_1(w)\alpha} \cdot (u_2)^{F_2(w)+h_2(w)\alpha}$
then Halt else go to Step 3.

Step 3: $s \leftarrow (u_1)^{P_1(w)} \cdot (u_2)^{P_2(w)}$

Step 4: $m \leftarrow e \otimes H'(s)$

Step 5: If the resulting plaintext is a 0^k conclude
that C is an encryption of w .

The security of this scheme relies on DDH and the collision resistant hash function as shown in the next Section 3.5.

3.5 Security of KRPEKS Against CKA

The K-Resilient IBE scheme [10] is an identity based encryption that is based on the Decisional Diffie Hellman problem (DDH). The security of such scheme is based on the difficulty of solving DDH and whether the hash functions used are collision resistant or not. In [1], the following theorem was proved.

Theorem. Let G be a group of prime order q . If DDH is hard in G then KRIBE is said to be IK-CCA secure. So for any adversary A attacking the anonymity of

KRIBE under a chosen ciphertext attack and making in total a $q_d(\cdot)$ decryption oracle queries, there exist a distinguisher D_A for DDH and an adversary C attacking the collision resistance of H such that

$$Adv_{KRIBE,A}^{IK-CCA}(K) \leq 2Adv_{G,D_A}^{DDH}(K) + 2Adv_{H,C}^{CR}(K) + (q_d(K) + 2)/(2^{k-3})$$

Since KRIBE is IK-CCA secure [1]. Therefore if an adversary knows two IDs ID_0, ID_1 and is given a ciphertext encrypted using one of the IDs. The adversary would not be able to guess which one was used unless the DDH is not hard or the hash function is not collision resistant. In this Section we show that since the PEKS was built from the KRIBE and KRIBE has key privacy notions then PEKS should logically be proved to be secure under a CKA.

Now if we compare both schemes the KRIBE in [10] and the KRPEKS we would notice that the key generation algorithm in the latter is the same as the setup in the former. The trapdoor is created the same way a secret key is created for an ID in the IBE scheme but instead of the IDs we have words. The encryption of IBE is equivalent to it for the PEKS but instead of a message, a k length zero string 0^k is encrypted. The testing of the existence of a keyword is done by using the same decryption algorithm of the IBE and checking whether the result is equal to 0^k . In other words if an adversary can not tell the difference between which ID was used to encrypt a given ciphertext. Then the same adversary would not know which word was used in creating ciphertext $C = PEKS(A_{pub}, W_b)$. The only difference between the two security notions is that instead of having a decryption oracle in proving IK-CCA in KRIBE we have a trapdoor oracle in the new PEKS. So someone can conclude that the advantage of guessing the right word depends on the DDH problem, the collision resistance of the hash function and Q_t where Q_t is the maximum number of trapdoor queries issued by the adversary.

3.6 Constructing K-Resilient SCF-PEKS Scheme

In [2] the authors constructed a SCF-PEKS using the same methodology used in the PEKS in [4]. In this Section we will try to build a SCF-PEKS using the KR-PEKS described in 3.4.

- *SCF – KRPEKS – CPG* (Common Parameter Generator) :
 - Step 1: Choose a group G and two generators g_1, g_2
 - Step 2: Choose random k
 - Step 3: Choose a random collision resistant hash function H (Section 2.2).
 - Step 4: Calculate the common parameter $cp = \langle G, g_1, g_2, H, k \rangle$
- *SCF – KRPEKS – SKG*(cp) (Server Key Generator) :
 - Step 1: Choose 6 random k degree polynomials, chosen over Z_q

$$P_1(x) = d_0 + d_1x + d_2x^2 + \dots + d_kx^k ; P_2(x) = d'_0 + d'_1x + d'_2x^2 + \dots + d'_kx^k$$

$$F_1(x) = a_0 + a_1x + a_2x^2 + \dots + a_kx^k ; F_2(x) = a'_0 + a'_1x + a'_2x^2 + \dots + a'_kx^k$$

$$h_1(x) = b_0 + b_1x + b_2x^2 + \dots + b_kx^k ; h_2(x) = b'_0 + b'_1x + b'_2x^2 + \dots + b'_kx^k$$

Step 2: For $0 \leq t \leq K$; Compute $A_t = g_1^{a_t} g_2^{a_t}, B_t = g_1^{b_t} g_2^{b_t}, D_t = g_1^{d_t} g_2^{d_t}$

Step 3: Assign $A_{priv_s} = \langle F_1, F_2, h_1, h_2, P_1, P_2 \rangle$

$$A_{pub_s} = \langle g_1, g_2, A_0, \dots, A_k, B_0, \dots, B_k, D_0, \dots, D_k, H \rangle$$

– *SCF – KRPEKS – RKG* (Receiver Key Generator) :

Step 1: Choose 6 random k degree polynomials, chosen over Z_q

$$\hat{P}_1(x) = \hat{d}_0 + \hat{d}_1 x + \hat{d}_2 x^2 + \dots + \hat{d}_k x^k ; \hat{P}_2(x) = \hat{d}'_0 + \hat{d}'_1 x + \hat{d}'_2 x^2 + \dots + \hat{d}'_k x^k$$

$$\hat{F}_1(x) = \hat{a}_0 + \hat{a}_1 x + \hat{a}_2 x^2 + \dots + \hat{a}_k x^k ; \hat{F}_2(x) = \hat{a}'_0 + \hat{a}'_1 x + \hat{a}'_2 x^2 + \dots + \hat{a}'_k x^k$$

$$\hat{h}_1(x) = \hat{b}_0 + \hat{b}_1 x + \hat{b}_2 x^2 + \dots + \hat{b}_k x^k ; \hat{h}_2(x) = \hat{b}'_0 + \hat{b}'_1 x + \hat{b}'_2 x^2 + \dots + \hat{b}'_k x^k$$

Step 2: For $0 \leq t \leq K$; Compute $\hat{A}_t = g_1^{\hat{a}_t} g_2^{\hat{a}_t}, \hat{B}_t = g_1^{\hat{b}_t} g_2^{\hat{b}_t}, \hat{D}_t = g_1^{\hat{d}_t} g_2^{\hat{d}_t}$

Step 3: Choose a random collision resistant hash function H' (Section 2.2).

Step 4: Assign $A_{priv_r} = \langle \hat{F}_1, \hat{F}_2, \hat{h}_1, \hat{h}_2, \hat{P}_1, \hat{P}_2 \rangle$

$$A_{pub_r} = \langle g_1, g_2, \hat{A}_0, \dots, \hat{A}_k, \hat{B}_0, \dots, \hat{B}_k, \hat{D}_0, \dots, \hat{D}_k, H, H' \rangle$$

– *SCF – KRPEKS*

Step 1: Choose a random $r_1 \in Z_q$

Step 2: Compute $u_1 = g_1^{r_1} ; u_2 = g_2^{r_1}$

Step 3: Calculate $A_w \leftarrow \prod_{t=0}^k A_t^{w^t} ; B_w \leftarrow \prod_{t=0}^k B_t^{w^t} ; D_w \leftarrow \prod_{t=0}^k D_t^{w^t}$
 $\hat{A}_w \leftarrow \prod_{t=0}^k \hat{A}_t^{w^t} ; \hat{B}_w \leftarrow \prod_{t=0}^k \hat{B}_t^{w^t} ; \hat{D}_w \leftarrow \prod_{t=0}^k \hat{D}_t^{w^t}$

Step 4: $s \leftarrow D_w^{r_1} \hat{D}_w^{r_1}$

Step 5: $e \leftarrow (0^k) \otimes H'(s)$

Step 6: $\alpha \leftarrow H(u_1, u_2, e)$

Step 7: $v_w \leftarrow ((A_w)(\hat{A}_w))^{r_1} \cdot ((B_w)(\hat{B}_w))^{r_1 \alpha}$

Step 8: $C \leftarrow \langle u_1, u_2, e, v_w \rangle$

– *SCF – KRPEKS – TG* (Trapdoor Generator) :

Calculate: $T_w = \langle \hat{F}_1(W), \hat{F}_2(W), \hat{h}_1(W), \hat{h}_2(W), \hat{P}_1(W), \hat{P}_2(W) \rangle$

– *SCF – KRPEKS – T* (Testing Algorithm) :

Step 1: $\alpha \leftarrow H(u_1, u_2, e)$

Step 2: Test if $v_w \neq (u_1)^{F_1(w)+h_1(w)\alpha+\hat{F}_1(w)+\hat{h}_1(w)\alpha}$
 $(u_2)^{F_2(w)+h_2(w)\alpha+\hat{F}_2(w)+\hat{h}_2(w)\alpha}$

then halt else go to next step

Step 3: $s \leftarrow (u_1)^{P_1(w)+\hat{P}_1(w)} \cdot (u_2)^{P_2(w)+\hat{P}_2(w)}$

Step 4: $m \leftarrow e \otimes H'(s)$

Step 5: If the resulting plaintext is a 0^k conclude that C is the encryption of w .

Notice that the testing part can not be done except by the server. Therefore, the trapdoor could be sent via public channels.

3.7 Constructing a K-Resilient PECK

In [8] [11] the authors constructed a PECK by adopting ideas from the PEKS in [4]. In this paper we will try to build a PECK scheme by adopting ideas from the KR-PEKS. The four algorithms that form this scheme are described as follows.

– KRPECK-KeyGen:

Step 1: Choose a group G of order q and two generators g_1, g_2

Step 2: Choose 6 random k degree polynomials, chosen over Z_q

$$P_1(x) = d_0 + d_1x + d_2x^2 + \dots + d_kx^k ; P_2(x) = d'_0 + d'_1x + d'_2x^2 + \dots + d'_kx^k$$

$$F_1(x) = a_0 + a_1x + a_2x^2 + \dots + a_kx^k ; F_2(x) = a'_0 + a'_1x + a'_2x^2 + \dots + a'_kx^k$$

$$h_1(x) = b_0 + b_1x + b_2x^2 + \dots + b_kx^k ; h_2(x) = b'_0 + b'_1x + b'_2x^2 + \dots + b'_kx^k$$

Step 2: For $0 \leq t \leq k$; Compute $A_t = g_1^{a_t} g_2^{a_t}, B_t = g_1^{b_t} g_2^{b_t}, D_t = g_1^{d_t} g_2^{d_t}$

Step 3: Choose two random numbers $s_0, s_1 \in Z_q$

Step 4: Calculate $S = g_1^{s_0} . g_2^{s_1}$

Step 5: Choose a random collision resistant hash function H (Section 2.2).

Step 6: Calculate: $A_{priv} = \langle F_1, F_2, h_1, h_2, P_1, P_2, s_0, s_1 \rangle$

$$A_{pub} = \langle g_1, g_2, A_0, \dots, A_k, B_0, \dots, B_k, D_0, \dots, D_k, S, H \rangle$$

– KRPECK:

Step 1: Choose a random $r_1 \in Z_q$

Step 2: Compute $u_1 = g_1^{r_1} ; u_2 = g_2^{r_1}$

Step 3: Calculate for every W_i where $1 \leq i \leq m$

$$A_{w_i} \leftarrow \prod_{t=0}^k A_t^{w_t^i} ; B_{w_i} \leftarrow \prod_{t=0}^k B_t^{w_t^i} ; D_{w_i} \leftarrow \prod_{t=0}^k D_t^{w_t^i}$$

Step 4: Calculate e_i where $1 \leq i \leq m$ and $e_i \leftarrow D_{w_i}^{r_1}$

Step 5: Calculate α_i where $1 \leq i \leq m$ and $\alpha_i \leftarrow H(u_1, u_2, e_i)$

Step 6: $v_{w_i} \leftarrow (A_{w_i})^{r_1} . (B_{w_i})^{r_1 \alpha_i}$

Step 7: $C \leftarrow \langle u_1, u_2, e_1, \dots, e_m, v_{w_1}, \dots, v_{w_m}, S^{r_1} \rangle$

– KRPECK-Trapdoors:

Step 1: Choose $\Omega_1, \dots, \Omega_t$ where t is the number of keywords you want to search for and Ω_i is the keyword in position I_i

Step 2: $T_1 = P_1(\Omega_1) + P_1(\Omega_2) + \dots + P_1(\Omega_t) + s_0$

Step 3: $T_2 = P_2(\Omega_1) + P_2(\Omega_2) + \dots + P_2(\Omega_t) + s_1$

Step 4: For $1 \leq j \leq t$ Compute $\alpha_j \leftarrow H(u_1, u_2, e_{I_j})$

Step 5: $T_3 = F_1(\Omega_1) + \dots + F_1(\Omega_t) + h_1(\Omega_1)\alpha_1 + \dots + h_1(\Omega_t)\alpha_t$

Step 6: $T_4 = F_2(\Omega_1) + \dots + F_2(\Omega_t) + h_2(\Omega_1)\alpha_1 + \dots + h_2(\Omega_t)\alpha_t$

Step 7: $T_Q = \langle T_1, T_2, T_3, T_4, I_1, \dots, I_t \rangle$

– KRPECK-Test:

Step 1: Test if $v_{w_{I_1}} . v_{w_{I_2}} \dots v_{w_{I_t}} \neq u_1^{T_3} . u_2^{T_4}$ then halt else do Step 2

Step 2: If $S^{r_1} . e_{I_1} . e_{I_2} . e_{I_3} \dots e_{I_t} = u_1^{T_1} . u_2^{T_2}$

then output “True” otherwise output false

If all the words exist, then definitely the condition of the if-statement in Step 2 in the testing algorithm will be true.

4 Conclusion

The main aim of this research was to have a PEKS that is secure under a standard model rather than the random oracle model only. To do so, the first step was finding an IBE scheme that has key privacy notions. Use of IBE with Weil pairing to build a PEKS was demonstrated in [5] and the scheme was secure under a

chosen keyword attack but under the random oracle only. The IBE suggested by Boneh and Boyen in [6] also was not useful in constructing a PEKS as shown in [2]. It was tempting to try to prove the K-resilient IBE [1] to have a notion of key privacy because it was shown in [3] that the Cramer-Shoup encryption is secure. The KRIBE adopted a lot of techniques from this encryption scheme and KRIBE was proved to be secure [1].

The new PEKS scheme was then used to construct a public key encryption with conjunctive keyword search and a public key encryption that does not need a secure channel.

However, the new PEKS scheme still has some drawbacks because of the limitations of the KRIBE scheme itself, where the number of malicious users is restricted to some value K . That is the number of trapdoors generated in the PEKS is limited to at most K . Nevertheless, that is not a serious problem where we could use a reasonably large K for email searching applications.

An additional concern lies in the basic formulation of the PEKS system. The idea of the sender, Bob having the sole power to decide which words to consider as keywords for the recipient, Alice, may not be as convenient in reality. In fact, Alice should have all influence on the email sorting and one solution would be for her to cache a set of criteria in form of queries. In that way, the emails categorized as ‘urgent’ would have greater possibility of being what she considers imperative to read.

References

1. Public key encryption with keyword search based on k-resilient ibe (full version).
2. J. Baek, R. Naini, and W. Susilo. Public key encryption with keyword search revisited. *Cryptology ePrint Archive*, Report 2005/191, 2005. <http://eprint.iacr.org/>.
3. M. Bellare, A. Boldyreva, A. Desai, and D. Pointcheval. Key-privacy in public-key cryptography. In *Advances in Cryptology - ASIACRYPT 2001*, volume 2248 of *Lecture Notes in Computer Science*, pages 566–582. Springer-Verlag, 2001.
4. D. Boneh, G. Crescenzo, R. Ostrovsky, and G. Persiano. Public-key encryption with keyword search. In C. Cachin, editor, *Proceedings of Eurocrypt 2004*, 2004. citeseer.ist.psu.edu/boneh04public.html.
5. D. Boneh and M. Franklin. Identity based encryption from the Weil pairing. *SIAM Journal on Computing*, 32(3):586–615, 2003.
6. Dan Boneh and Xavier Boyen. Short signatures without random oracles. *Cryptology ePrint Archive*, Report 2004/171, 2004. <http://eprint.iacr.org/>.
7. R. Cramer and V. Shoup. A practical public key cryptosystem provably secure against adaptive chosen ciphertext attack. In *Advances in Cryptology - CRYPTO '98*, volume 1462 of *Lecture Notes in Computer Science*, pages 13–25. Springer-Verlag, 1998.
8. J. Cha D. Park and P. Lee. Searchable keyword-based encryption. *Cryptology ePrint Archive*, Report 2005/367, 2005. <http://eprint.iacr.org/>.
9. R. Hayashi and K. Tanaka. Elgamal and cramer shoup variants with anonymity using different groups extended abstract. C-200, 2004. <http://www.is.titech.ac.jp/research/research-report/C/>.

10. S. Heng and K. Kurosawa. K-resilient identity-based encryption in the standard model. In *Topics in Cryptology CT-RSA 2004*, volume 2964, pages 67–80. Springer-Verlag, 2004.
11. D. Park, K. Kim, and P. Lee. Public key encryption with conjunctive field keyword search. *Lecture Notes in Computer Science*, 3325:73–86, 2005.
12. A. Shamir. Identity-based cryptosystems and signature schemes. In *Advances in Cryptology - CRYPTO '84*, volume 0193 of *Lecture Notes in Computer Science*, pages 47–53. Springer-Verlag, 1984.

A Generic Construction of Secure Signatures Without Random Oracles

Jin Li¹, Yuen-Yan Chan², and Yanming Wang^{1,3}

¹ School of Mathematics and Computational Science,
Sun Yat-Sen University,
Guangzhou, 510275, P.R. China
sysjinli@yahoo.com.cn

² Department of Information Engineering,
Chinese University of Hong Kong,
Shatin, N.T., Hong Kong
yychan@ie.cuhk.edu.hk

³ Lingnan College, Sun Yat-Sen University,
Guangzhou, 510275, P.R. China
stswym@zsu.edu.cn

Abstract. We show how to construct an existentially unforgeable secure signature scheme from any scheme satisfies only a weak notion of security in the standard model. This construction method combines a weakly secure signature and a one-time signature. However, key generation of the resulted fully secure signature is the same as the key generation of weak signature. Therefore the length of the public key in our fully secure signature is independent of that of the one-time signature. Our conversion from a weakly secure signature scheme to an existentially unforgeable secure signature scheme is simple, efficient and provably secure in the standard model (that is, security of the resulting scheme does not rely on the random oracle model). Our results yield a new construction of existentially unforgeable secure signature in the standard model. Furthermore, we show two efficient instantiations without random oracles converted from two previous weakly secure signature schemes.

Keywords: Signature, Standard Model, Weak Chosen Message Attack.

1 Introduction

Digital signature is a central cryptographic primitive and it is one of the fundamental building blocks in cryptographic protocols. After Goldwasser, Micali and Rivest [17] formally defined the standard notion of security for digital signatures, namely existentially unforgeability under an adaptive chosen message attack, they have been studied widely. Since then, there have been many attempts to design practical and provably secure signature schemes.

Provably secure signature schemes can be constructed from the most basic cryptographic primitive, one-way function [2,19,20]. Most of them were constructed from authentication trees and one-time signatures. They are very not

practical for the public key of these schemes grows with the number of messages to be signed, which is often the case with cryptographic schemes designed from elementary blocks. Over the years several signature schemes have been proposed based on stronger complexity assumptions. The most efficient schemes provably secure in the standard model are based on the Strong-RSA assumption [11,16] and q -strong Diffie-Hellman (q -SDH) assumption [5]. Different from standard model, the random oracle model was introduced by Bellare and Rogaway [3], many of such schemes [3,4,7,23] were also constructed in this kind of model. However, security proofs in random oracle model can only be heuristic as Canetti, Goldreich and Halevi [16] showed that behaving like a random oracle is not a property that can be realized in general; and that security proofs in the random oracle model do not always imply the security of the actual scheme in the real world. Therefore, signatures provably secure in the standard model attract a great interest.

1.1 Related Work

A weaker notion of security for signatures, proposed in [5], requires the adversary to submit all signature queries before the system parameters are published. Security of this kind is called existential unforgeability under a weak chosen message attack [5]. More detailed definition will be given in Section 2.

It has been shown [5] that a weakly secure signature scheme can be constructed based on the q -SDH assumption in the standard model. Gennaro, Halevi and Rabin proposed a secure signature scheme without random oracles, however, under the assumptions of Strong-RSA and that the hash function is a collision-free, division intractable, and is a non-standard randomness-finding oracle. In fact, if without the assumption of the randomness-finding oracle, the scheme in [16] can only be proven secure in the weak chosen message attack model. In order to obtain fully secure signatures, both [5,16] used similar chameleon hash method to convert the weakly secure signature schemes to fully secure signature schemes.

1.2 Our Contribution

We present here a new construction of existentially unforgeable secure signature scheme based on any signature schemes satisfying the relatively weak notion of security and secure one-time signature (OTS). Our conversion from a weak secure signature scheme to an existentially unforgeable secure signature scheme is simple, efficient and provably secure in the standard model. We also show two efficient instantiations without random oracles converted from two previous weak secure signature schemes. Compared to previous provably secure signature in the standard model, length of public key of the new signature is very short. Meanwhile, the only online computation in signature generation is the one that required by the OTS , which makes our scheme very efficient.

1.3 Organization

Our paper is organized in the following way. Section 2 provides security definitions. Our construction is presented in Section 3. Security theorems and the

corresponding proofs are given in Section 4. Two instantiations are described in Section 5. The paper is concluded in Section 6.

2 Definitions

A signature scheme is made up of three algorithms, **Gen**, **Sign**, and **Verify**, which are for generating keys, signing, and verifying signatures respectively. The standard notion of security for a signature scheme is called existential unforgeability under a chosen message attack [17], which is defined through the following game between a challenger \mathcal{C} and an adversary \mathcal{A} :

1. \mathcal{C} runs $\text{Gen}(1^k)$ and obtains a public key PK and secret key SK . The public key PK is sent to \mathcal{A} .
2. \mathcal{A} requests signatures on at most q_S messages m_i adaptively for $i = 1, \dots, q_S$, \mathcal{C} returns the corresponding signature σ_i which is obtained by running algorithm **Sign**.
3. Finally, \mathcal{A} outputs (m^*, σ^*) , where m^* is a message, and σ^* is a signature, such that m^* are not equal to the inputs of any query to **Sign**. \mathcal{A} wins the game if σ is a valid signature of m^* .

A signature is secure if \mathcal{A} cannot output a valid forged signature after the above game. The security definition of \mathcal{OTS} is the same as signatures, except that the attacker is restricted to query the signing oracle for only one time, that is, $q_S = 1$.

A slightly stronger notion of security, called strong existential unforgeability [1], which is also defined using the above game between a challenger \mathcal{C} and an adversary \mathcal{A} , except the definition that \mathcal{A} wins the game is \mathcal{A} can output a pair (m^*, σ^*) such that (m^*, σ^*) is not any of (m_i, σ_i) and $\text{Verify}_{PK}(m^*, \sigma^*) = 1$.

We also review the definition of a weaker notion of security, namely existential unforgeability under a weak chosen message attack defined by Boneh *et. al.* [5]. The difference from [17] is that here the adversary is required to submit all messages for signature queries before the public parameters are published. Similarly, we can also define the notion called strong existential unforgeability under the weak chosen message attack such that the adversary is required to submit all messages for signature queries before the public parameters are published. We say \mathcal{A} wins the game if it outputs a pair (m^*, σ^*) such that (m^*, σ^*) is not any of (m_i, σ_i) and $\text{Verify}_{PK}(m^*, \sigma^*) = 1$.

3 Fully Secure Signature from Weakly Secure Signature

It has been shown [5,16] that a weakly secure signature scheme can be constructed based on the strong-RSA and q -SDH assumption in the standard model. In fact, both [5,16] use the chameleon hash method to convert a weakly secure signature scheme to a fully secure signature scheme. In our construction of existentially unforgeable secure signature, we require only a signature scheme satisfying this weaker notion of security.

The conversion of any such weak secure signature scheme to an existentially unforgeable secure signature scheme is described as follows. The public key PK of the new scheme is simply the public key of the weak secure signature scheme, and the secret key SK is the corresponding secret key. To sign a message, the signer first generates a key-pair (vk, sk) for \mathcal{OTS} . The signer then signs vk using SK , and signs the message with the secret key sk . The resulting signature consists of vk , a signature on vk signed with SK , and signature on message m signed with sk . To verify the signature on message m , the receiver first verifies the signature on vk with respect to PK and the signature on m with respect to vk .

Given a signature scheme $\mathcal{S}' = (\text{Gen}'; \text{Sign}'; \text{Verify}')$ secure against weak chosen message attack, we construct an existentially unforgeable secure signature scheme $\mathcal{S} = (\text{Gen}; \text{Sign}; \text{Verify})$ against adaptively chosen message attack. In the construction, we use a one-time signature scheme $\mathcal{OTS} = (\text{OGen}; \text{OSign}; \text{OVerify})$, where OGen , OSign , and OVerify are the key generation algorithm, signing algorithm, and signature verification algorithm respectively.

The construction of \mathcal{S} proceeds as follows:

1. **Gen:** On input of the security parameter 1^k , invoke $\text{Gen}'(1^k)$ and obtain $(PK, SK) \leftarrow \text{Gen}'(1^k)$. Output \mathcal{S} 's public key PK and secret key SK (In fact, $\text{Gen} = \text{Gen}'$).
2. **Sign:** To sign message m , the signer first invokes $\text{OGen}(1^k)$ to obtain the one-time signature key pair $(vk, sk) \leftarrow \text{OGen}(1^k)$. The signer then invokes algorithms $\text{Sign}'_{SK}(vk)$ and $\text{OSign}_{sk}(m)$. Output $\sigma = (A, B, C)$ as the signature, where $A = \text{Sign}'_{SK}(vk)$, $B = \text{OSign}_{sk}(m)$, $C = vk$.
3. **Verify:** On input verifying key PK , message m , and $\sigma = (A, B, C)$, output 1 if and only if $\text{Verify}'_{PK}(A, C) = 1$ and $\text{OVerify}_C(m, B) = 1$.

Key generation of the resulted fully secure signature is the same as the key generation of the weak signature. Therefore the length of the public key in the fully secure signature is independent of that of the one-time signature scheme. Our conversion from a weakly secure signature scheme to an existentially unforgeable secure signature scheme is simple and efficient in the standard model. In signature generation phase, $\text{Sign}'_{SK}(vk)$ can be pre-computed by the signer. Therefore the online computation in signature generation is only the computation of the one-time signature \mathcal{OTS} , which is very efficient. Computation required by the verification in our scheme is the same as the computation required by the verification of \mathcal{S}' and \mathcal{OTS} .

4 Security Results

We first give some intuition as to why \mathcal{S} is secure against adaptively chosen message attack. Given only weakly secure signature \mathcal{S}' and one-time signature \mathcal{OTS} , the simulator can answer the adaptively chosen signature queries from the adversary because the chosen one-time key is independent of the message chosen by the adversary, which implies that the one-time public keys can be sent to \mathcal{S}' for signatures before messages are given, and then \mathcal{OTS} is used to sign messages by the adversary.

Let $\sigma_i = (A_i, B_i, C_i)$ be the queried signature and let $\sigma^* = (A^*, B^*, C^*)$ be the forged signature on a new message m^* output by the adversary. On one hand, if $C^* \neq C_i$ for $i = 1, \dots, q_S$, then it implies that the \mathcal{S}' is insecure under weakly chosen message attack. On the other hand, if $C^* = C_i$ for some signature output by the simulator, then B^* is another valid signature with respect to the one-time key C^* . That is, the adversary breaks the one-time signature scheme. Therefore under the assumption that \mathcal{S}' is secure under weakly chosen message attack and that \mathcal{OTS} is a secure one-time signature, the signature \mathcal{S} is existentially unforgeable under adaptively chosen message attack.

Next we formally prove the security of the signature scheme \mathcal{S} .

Theorem 1. *If \mathcal{S}' is a signature scheme which is existentially unforgeable secure against weak chosen message attack, and \mathcal{OTS} is an unforgeable one-time signature scheme, then \mathcal{S} is a signature scheme which is existentially unforgeable secure against adaptive chosen message attack.*

Proof. Given any adversary \mathcal{A} attacking \mathcal{S} in an adaptive chosen message attack, we construct an adversary \mathcal{A}' breaking \mathcal{S}' in a weak chosen message attack or breaking \mathcal{OTS} . After giving public key PK of \mathcal{S} , \mathcal{A} queries the signing oracle of \mathcal{S} on messages m_i adaptively and gets q_S signatures $\sigma_i = (A_i, B_i, C_i)$ for $1 \leq i \leq q_S$. After the signature queries, \mathcal{A} outputs a forged signature on a new m^* as $\sigma^* = (A^*, B^*, C^*)$.

There are two types of forgeries while the reduction works differently for each type. Therefore, initially \mathcal{A}' may choose a random bit $b_{code} \in \{1, 2\}$ that indicates its guess for the type of forger that \mathcal{A} will emulate. The simulation proceeds differently for each b_{code} :

Type 1 Forgery. $C^* \neq C_i$ for $1 \leq i \leq q_S$.

Algorithm \mathcal{A}' first picks a random bit b_{code} . If $b_{code} = 1$, we construct an algorithm \mathcal{A}' to break \mathcal{S}' . \mathcal{A}' first invokes $\text{OGen}(1^k)$ and gets q_S key pairs $(vk_i, sk_i) \leftarrow \text{OGen}(1^k)$ for \mathcal{OTS} (assume \mathcal{A} makes at most q_S queries to the signing oracle), and sends the q_S values vk_i , for $1 \leq i \leq q_S$, to challenger for signature queries of \mathcal{S}' before the parameters of \mathcal{S}' are published. Then \mathcal{A}' gets public key PK of \mathcal{S}' and q_S signatures σ'_i on the q_S messages vk_i for $1 \leq i \leq q_S$. Then \mathcal{A}' sends the public key PK to the adversary \mathcal{A} as the public key of \mathcal{S} . \mathcal{A} then queries the signing oracle of \mathcal{S} on messages m_i adaptively for $1 \leq i \leq q_S$. \mathcal{A}' answers the signature query as follows: $\sigma_i = (A_i, B_i, C_i)$, where $A_i = \sigma'_i$ from the challenger, $B_i = \text{OSign}_{sk_i}(m_i)$, and $C_i = vk_i$. After the signature queries, \mathcal{A} outputs a forged signature on a new message m^* as $\sigma^* = (A^*, B^*, C^*)$. Because $C^* \neq C_i$ for $1 \leq i \leq q_S$, \mathcal{A}' can output a forged \mathcal{S}' signature as $\sigma = A^*$ on a new message C^* and break the signature scheme \mathcal{S}' .

Type 2 Forgery. $C^* = C_i$ for some i , $1 \leq i \leq q_S$.

If $b_{code} = 2$, we construct an algorithm \mathcal{A}' to break \mathcal{OTS} . \mathcal{A}' is given vk^* from the challenger as the challenge public key for \mathcal{OTS} . Then it randomly generates $(SK, PK) \leftarrow \text{Gen}'(1^k)$ of \mathcal{S}' . \mathcal{A}' then gets the key pair (SK, PK) of \mathcal{S} and sends the public key PK to \mathcal{A} . \mathcal{A}' also chooses a random $\kappa \in [1, q_S]$ and keeps

it secret. Next \mathcal{A} queries the signing oracle of \mathcal{S} on messages m_i adaptively for $1 \leq i \leq q_S$. \mathcal{A}' answers the signature query as follows: if $i \neq \kappa$, \mathcal{A}' computes one-time key pair (sk_i, vk_i) , returns the signature as $\sigma_i=(A_i, B_i, C_i)$, where $A_i = \text{Sign}'_{SK}(vk_i)$, $B_i = \text{OSign}_{sk_i}(m_i)$, $C_i = vk_i$. Otherwise, if $i = \kappa$, \mathcal{A}' sends m_i to the challenger for one-time signature with respect to public key vk^* and gets the one-time signature B_i on message m_i . Then \mathcal{A}' answers the signature query as $\sigma_i=(A_i, B_i, C_i)$, where $B_i = \text{Sign}'_{SK}(vk^*)$ and $C_i = vk^*$. After the signature queries, \mathcal{A} outputs a forged signature on a new message m^* as $\sigma^* = (A^*, B^*, C^*)$, where $C^* = C_i$ for some $1 \leq i \leq q_S$. If $i \neq \kappa$, \mathcal{A}' aborts and fails. If $i = \kappa$ (with success probability $\frac{1}{q_S}$), then $C^* = vk^*$. Meanwhile, $B^* \neq B_i$ because $m^* \neq m_i$. This implies \mathcal{A}' can output a forged one-time signature B^* on a new message m^* with respect to vk^* and break the one-time signature scheme \mathcal{OTS} .

Therefore \mathcal{S} is existential unforgeable secure against adaptively chosen message attack under the assumptions that \mathcal{S}' is a weak secure signature scheme and \mathcal{OTS} is a unforgeable one-time signature scheme.

We can also get the following result similarly.

Theorem 2. *If \mathcal{S}' is a signature scheme which is strongly existentially unforgeable secure against weak chosen message attack and \mathcal{OTS} is an strongly unforgeable one-time signature scheme, then \mathcal{S} is a signature scheme which is strongly existentially unforgeable secure against adaptive chosen message attack.*

Proof. Similar to the proof of Theorem 1.

5 Efficient Instantiations

5.1 Fully Secure Signature from [16]

Gennaro, Halevi and Rabin [16] proposed a secure signature scheme without random oracle, however, under the assumption of Strong-RSA and that hash function is collision and division intractable, and it has a non-standard randomness-finding assumption. In fact, without the assumption of randomness-finding oracle, the scheme in [16] can only be proven secure in the weak chosen message attack model.

Definition 1 (Strong-RSA Assumption). *Given a randomly chosen RSA modulus n , and a random element $s \in \mathbb{Z}_n^*$, it is infeasible to find a pair (e, r) with $e > 1$ such that $r^e = s \pmod n$.*

Next, we show the converted secure signature under adaptively chosen message attack from weak secure signature and $\mathcal{OTS} = (\text{OGen}; \text{OSign}; \text{OVerify})$. Define a hash function H which is collision and division intractable satisfies $H : (0, 1)^* \rightarrow \mathbb{Z}_n^*$. The fully secure signature scheme is described below:

1. **Gen:** Pick two safe primes p and q , compute $n = pq$ as RSA modulus, select $s \in \mathbb{Z}_n^*$. The public key is $PK=(n, s)$ and the secret key is $SK=(p, q)$.
2. **Sign:** To sign a message m , invoke $\text{OGen}(1^k)$ and obtain key pair $(vk, sk) \leftarrow \text{OGen}(1^k)$ of \mathcal{OTS} . Output the signature as $\sigma = (A, B, C)$, where $A = s^{\frac{1}{H(vk)}} \pmod n$, $B = \text{OSign}_{sk}(m)$, $C = vk$.

3. **Verify:** On input verification key (n, s) , message m , and $\sigma = (A, B, C)$, output 1 if and only if $A^{H(C)} = s \pmod n$ and $\text{OVerify}_C(m, B) = 1$. Otherwise, output 0.

We first show the underlying weak secure signature scheme $\mathcal{S}' = (\text{Gen}' ; \text{Sign}' ; \text{Verify}')$ briefly, which can be easily provable based on Strong-RSA assumption under weak chosen message attack. Gen' outputs the public key (n, s) and secret key (p, q) . On input a message m , the signer computes the signature as $\sigma = s^{\frac{1}{H(m)}} \pmod n$. On input m and σ , Verify' outputs 1 if $\sigma^{H(m)} = s \pmod n$.

It requires one exponentiation in \mathbb{Z}_n^* and a one-time signature computation in the signature generation. In fact, the value A and C can be pre-computed. So the only online computation of Sign is the computation of the one-time signature, which is more efficient than [11,16]. And it does not require the non-standard randomness-finding assumption compared with [16].

The converted signature scheme can also be provable from Theorem 1 for the underlying signature is weak secure under Strong-RSA assumption and that \mathcal{OTS} is secure.

5.2 Fully Secure Signature from [5]

Preliminary. Let \mathbb{G}_1 be a multiplicative group generated by g , whose order is a prime p , and \mathbb{G}_2 also be a multiplicative group with the same order p . Let $e : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$ be a map with the following properties: bilinearity, non-degeneracy and computability. As shown in [5,7], such non-degenerate bilinear map over cyclic groups can be obtained from the Weil or the Tate pairing over algebraic curves.

Definition 2 (q -Strong Diffie-Hellman Assumption). *The q -SDH assumption in group \mathbb{G}_1 is defined as follows: given a $(q+1)$ -tuple $(g, g^x, g^{x^2}, \dots, g^{x^q}) \in (\mathbb{G}_1)^{q+1}$ as input, output a pair $(c, g^{1/(x+c)})$ is hard, where $c \in \mathbb{Z}_p^*$.*

Next, we describe the converted secure signature under adaptively chosen message attack from the weak secure signature [4] and \mathcal{OTS} . Let $(\mathbb{G}_1, \mathbb{G}_2)$ be bilinear groups where the order of both \mathbb{G}_1 and \mathbb{G}_2 is p . As usual, g is a generator of \mathbb{G}_1 . $\mathcal{OTS} = (\text{OGen} ; \text{OSign} ; \text{OVerify})$ is a one-time signature. Meanwhile, define a collision-resistant hash function $H : \{0, 1\}^* \rightarrow \mathbb{Z}_p^*$.

1. **Gen:** Pick $x \in \mathbb{Z}_p^*$, compute $y = g^x$. The public key is $PK=(g, y)$ and the secret key is $SK=x$.
2. **Sign:** Given message $m \in \mathbb{Z}_p^*$, the signer first invokes $\text{OGen}(1^k)$ and obtains key pair $(vk, sk) \leftarrow \text{OGen}(1^k)$. Output the signature on m as $\sigma = (A, B, C)$, where $A = g^{\frac{1}{x+H(vk)}}$, $B = \text{OSign}_{sk}(m)$, $C = vk$.
3. **Verify:** On input verification key y , message m , and the signature $\sigma = (A, B, C)$, output 1 if and only if $e(y \cdot g^{H(C)}, A) = e(g, g)$ and $\text{OVerify}_C(m, B) = 1$. Otherwise, output 0.

Notice that the user's public key consists only one group element y in \mathbb{G}_1 . So the length of the public key is even shorter than that in [5]. It requires one point scalar multiplication in \mathbb{G}_1 and one one-time signature computation in signature generation. In fact, the value A and C can be pre-computed. So the online computation of **Sign** is only the computation of the one-time signature, which is very efficient compared ordinary signature schemes. Verification only requires two pairing computations, one point scalar multiplication in \mathbb{G}_1 , and an *OTS* verification, which is also very efficient. The only disadvantage of this signature scheme is that length of the signature is longer than that in [5].

The signature scheme can be proven to be secure from Theorem 1 for the underlying signature is weak secure under q -SDH assumption and that *OTS* is secure.

6 Conclusion

We have shown how to construct an existentially unforgeable secure signature scheme from any scheme satisfies only a weak notion of security which is known to be achievable in the standard model. The new method of designing secure signature is quite different from known methods. Moreover, the resulting signature schemes are simple, efficient, and provably secure in the standard model. Designing existentially unforgeable secure signature scheme under adaptively chosen message attack, then, can be reduced to the design of secure signature under only weak chosen message attack. Moreover, we have presented two concrete signature schemes without random oracles converted from two previous weak secure signature schemes [5,16]. The lengths of the public key of the new signature in both schemes are very short. Meanwhile, the new signature scheme converted from [16] does not rely on the non-standard randomness-finding oracle, which has been used in [16]. Furthermore, in both schemes, the only online computation required in signature generation is the computation of the one-time signature, which makes our schemes very efficient.

Acknowledgements

We thank Professor Victor K. Wei for valuable discussions, and acknowledgement to Hong Kong Research Grant Council's Earmarked Grants 4232-03E and 4328-02E for sponsorship, as well as National Natural Science Foundation of China 10571181 for financial support. Part of the work of the first author was done while visiting Chinese University of Hong Kong.

References

1. J.H. An, Y. Dodis, and T. Rabin. *On the security of joint signature and encryption*. In Proceedings of Eurocrypt 2002, volume 23-32 of LNCS. Springer-Verlag, 2002.
2. M. Bellare and S. Micali. *How to sign given any trapdoor function*. J. of the ACM 39,1992, pp. 214-233.

3. M. Bellare and P. Rogaway. *Random oracle are practical: A paradigm for designing efficient protocols*. In Proceedings of the First ACM Conference on Computer and Communications Security, pages 62-73, 1993.
4. M. Bellare and P. Rogaway. *The exact security of digital signatures: How to sign with RSA and Rabin*. In Ueli Maurer, editor, Proceedings of Eurocrypt'96, volume 1070 of LNCS, pages 399-416. Springer-Verlag, 1996.
5. D. Boneh and X. Boyen. *Short signatures without random oracles*. Proc. of Eurocrypt'04, LNCS 3027, pp. 56-73, Springer-Verlag, 2004.
6. D. Boneh, X. Boyen, and H. Shacham. *Short group signatures*. In Proceedings of Crypto 2004, LNCS 3152, pp.41-55, Springer-Verlag, 2004.
7. D. Boneh, B. Lynn, and H. Shacham. *Short signatures from the Weil pairing*. In Proceedings of Asiacrypt 2001, volume 2248 of LNCS, pages 514-532. Springer-Verlag, 2001.
8. R. Canetti, O. Goldreich and S. Halevi. *The Random Oracle Methodology, Revisited*. STOC'98, ACM, pp. 207-221, 1998.
9. J.-S. Coron and D. Naccache. *Security analysis of the Gennaro-Halevi- Rabin signature scheme*. In Proceedings of Eurocrypt 2000, pages 91-101, 2000.
10. J.-S. Coron. *On the exact security of full domain hash*. In Proceedings of Crypto 2000, volume 1880 of LNCS, pages 22-35. Springer-Verlag, 2000.
11. R. Cramer and V. Shoup. *Signature schemes based on the strong RSA assumption*. ACM TISSEC, 3(3):161-185, 2000. Extended abstract in Proc. 6th ACM CCS, 1999.
12. C. Dwork and M. Naor. *An efficient existentially unforgeable signature scheme and its applications*. In J. of Cryptology, 11(3), Summer 1998, pp. 187-208.
13. T. Elgamal, *A public key cryptosystem and a signature scheme based on discrete logarithms*, IEEE Trans. Info. Theory, IT-31(4), 1985, pp. 469-472.
14. S. Even, O. Goldreich, and S. Micali. *On-line/Off-line digital signatures*, Journal of Cryptology, vol 9, 1996, pp. 35-67.
15. A. Fiat and A. Shamir, *How to prove yourself*, advances of Cryptology, Crypto'86, LNCS 263, Springer-Verlag, 1987, pp. 641-654.
16. R. Gennaro, S. Halevi, and T. Rabin. *Secure hash-and-sign signatures without the random oracle*. In Proceedings of Eurocrypt 1999, LNCS, pages 123-139. Springer-Verlag, 1999.
17. S. Goldwasser, S. Micali, and R. Rivest. *A digital signature scheme secure against adaptive chosen-message attacks*. SIAM J. Computing, 17(2):281-308, 1988.
18. H. Krawczyk and T. Rabin. *Chameleon signatures*. In Proceedings of NDSS 2000. Internet Society, 2000. <http://eprint.iacr.org/1998/010/>.
19. Leslie Lamport. *Constructing digital signatures from a one way function*. Technical Report CSL-98, SRI International, October 1979.
20. A. Perrig. *The BiBa one-time signature and broadcast authentication protocol*. In Eighth ACM Conference on Computer and Communication Security, pages 28-37. ACM, 2001.
21. D. Pointcheval and J. Stern, *Security arguments for digital signatures and blind signatures*, Journal of Cryptology, Vol.13, No.3, pp.361-396, 2000.
22. R. Rivest, A. Shamir and L. Adleman, *A method for obtaining digital signature and public key cryptosystems*, Comm. of ACM, 21, 1978, pp. 120-126.
23. F. Zhang, R. Safavi-Naini, and W. Susilo, *An efficient signature scheme from bilinear pairings and its applications*, In Proceedings of PKC 2004, LNCS 2947, pp. 277-290, Springer-Verlag, 2004.

A Separation Between Selective and Full-Identity Security Notions for Identity-Based Encryption

David Galindo

Institute for Computing and Information Sciences, Radboud University Nijmegen,
P.O. Box 9010, 6500 GL, Nijmegen, The Netherlands
d.galindo@cs.ru.nl
<http://cs.ru.nl/~dgalindo>

Abstract. Identity-based encryption has attracted a lot of attention since the publication of the scheme by Boneh and Franklin. In this work we compare the two adversarial models previously considered in the literature, namely the full and selective-identity models. Remarkably, we show that the strongest security level with respect to selective-identity attacks (i.e. chosen-ciphertext security) fails to imply the weakest full-identity security level (i.e. one-wayness). In addition, an analogous result for the related primitive of tag-based encryption is presented.

Keywords: Foundations, identity-based encryption, tag-based encryption.

1 Introduction

IDENTITY-BASED ENCRYPTION. The concept of identity-based encryption (IBE) was proposed by Shamir in [Sha85], aimed at simplifying certificate management in e-mail related systems. The idea is that an arbitrary string such as an e-mail address or a telephone number could serve as a public key for an encryption scheme. Once a user U receives a communication encrypted using its identity id_U , the user authenticates itself to a Key Generation Center (KGC) from which it obtains the corresponding private key d_U .

The problem was not satisfactorily solved until the work by Boneh and Franklin [BF03]. They proposed formal security notions for IBE systems and designed a secure IBE scheme. Since then, IBE has attracted a lot of attention, and a large number of IBE schemes and related systems have been proposed.

The first IBE security notions were proposed in [BF03]. These new notions were inspired on the existing traditional public key encryption (PKE) definitions [RS92, NY90, GM84], with the novelties that the adversary, regardless of the attack model, is given access to an *extraction oracle*, which on input an identity id outputs the corresponding private key. Moreover, the adversary selects the identity id on which it wants to be challenged, so that now the challenge consists

of a pair (id, c) , where c denotes a ciphertext encrypted under identity id . In [BF03] the adversary is allowed to adaptively select id , probably depending on the information received so far (such as decryption keys), while in [CHK04] the adversary must commit ahead of time to the challenge identity. The latter model is referred to as *selective-identity* attack (sID), while the original model is called *full-identity* scenario (ID). With respect to IBE goals, only one-wayness and indistinguishability were defined in [BF03]. Thus, the security definitions mostly considered up to now in the literature are: OW-ID-CPA, IND-ID-CPA, IND-sID-CPA, IND-ID-CCA, IND-sID-CCA.

TAG-BASED ENCRYPTION. This notion was introduced by MacKenzie, Reiter and Yang in [MRY04]. Informally, in a tag-based encryption (TBE) scheme, the encryption and decryption operations take an additional tag. Such a tag is a binary string of appropriate length, and need not have any particular internal structure. In [MRY04] it is shown how to build chosen-ciphertext secure PKE schemes from (adaptive-tag) chosen-ciphertext secure TBE and one-time signature schemes. Interestingly, Kiltz [Kil06] has proposed the notion of selective-tag chosen-ciphertext security and has shown that a TBE scheme secure in this new sense suffices to obtain secure PKE schemes with the transformation in [MRY04].

OUR CONTRIBUTION. In this work we show that the strongest security level with respect to sID attacks (i.e. chosen-ciphertext security) do not even imply the weakest ID security level (i.e. one-wayness). Previous to our work, selective security was considered just a slightly weaker security model [BB04], but it turns out that selective-identity security is a strictly weaker security requirement than full-identity security. Notwithstanding, sID security suffices for other purposes, for instance for building chosen-ciphertext secure PKE schemes [CHK04], and there exists an efficient generic transformation in the Random Oracle Model [BR93] from sID security to ID security [BB04].

Additionally, we show that selective-tag security is a strictly weaker requirement than adaptive-tag security. Although a TBE scheme can be seen as some sort of “flattened” IBE scheme, that is, an IBE scheme without private key extraction operation, it does not seem possible to apply our IBE separation result to TBE schemes. Therefore we have used a different technique in the TBE case.

2 Definitions for Identity-Based Encryption

We start by fixing some notation and recalling basic concepts.

Algorithmic Notation. Assigning a value a to a variable x will be in general denoted by $x \leftarrow a$. If A is a non-empty set, then $x \leftarrow A$ denotes that x has been uniformly chosen in A . If D is a probability distribution over A , then $x \leftarrow D$ means that x has been chosen in A by sampling the distribution D . Finally, if \mathcal{A} is an algorithm, $x \leftarrow \mathcal{A}$ means that \mathcal{A} has been executed on some specified input and its output has been assigned to the variable x .

Negligible Functions. The class of negligible functions on a parameter $\ell \in \mathbb{Z}^+$ is the set of the functions $\epsilon : \mathbb{Z}^+ \rightarrow \mathbb{R}^+$ such that, for any polynomial $p \in \mathbb{R}[\ell]$, there exist $M \in \mathbb{R}^+$ such that $\epsilon(\ell) < \frac{M}{p(\ell)}$ for all $\ell \in \mathbb{Z}^+$.

Definition 1. An *identity-based encryption scheme* handling identities of length ℓ (where ℓ is a polynomially-bounded function) is specified by four probabilistic polynomial time (PPT) algorithms:

Setup. IBE.Gen takes a security parameter 1^k and returns the master public key PK and master secret key SK . The master public key PK include the security parameter 1^k and $\ell(k)$; as well as the description of sets \mathcal{M}, \mathcal{C} , which denote the set of messages and ciphertexts respectively.

Extract. IBE.Ext takes as inputs SK and $id \in \{0, 1\}^\ell$; it outputs the private key d_{id} corresponding to the identity id .

Encrypt. IBE.Enc takes as inputs PK , an identity $id \in \{0, 1\}^\ell$ and $m \in \mathcal{M}$. It returns a ciphertext $c \in \mathcal{C}$.

Decrypt. IBE.Dec takes as inputs a private key d_{id} and $c \in \mathcal{C}$, and it returns $m \in \mathcal{M}$ or reject when c is not a legitimate ciphertext. For the sake of consistency, these algorithms must satisfy $\text{IBE.Dec}(PK, d_{id}, c) = m$ for all $id \in \{0, 1\}^\ell$, $m \in \mathcal{M}$, where $c = \text{IBE.Enc}(PK, id, m)$ and $d_{id} = \text{IBE.Ext}(SK, id)$.

2.1 One-Wayness for IBE

In the following, the notion of one-wayness against full-identity chosen-plaintext attacks (referred to as OW-ID-CPA in the following definition) is recalled.

Definition 2. Let $\Pi = (\text{IBE.Gen}, \text{IBE.Ext}, \text{IBE.Enc}, \text{IBE.Dec})$ be an IBE scheme, and let $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ be any 2-tuple of PPT oracle algorithms. We say Π is OW-ID-CPA secure if for any 2-tuple of PPT oracle algorithms \mathcal{A} and any polynomially-bounded function $\ell(\cdot)$, the advantage in the following game is negligible in the security parameter 1^k :

1. IBE.Gen(1^k) outputs (PK, SK) . The adversary is given PK .
2. The adversary \mathcal{A}_1 may ask polynomially-many queries to an oracle $\text{IBE.Ext}(SK, \cdot)$ and finally outputs a challenge identity id^* from which it does not know the private key.
3. A message m is randomly chosen in \mathcal{M} and the adversary is given a challenge ciphertext $c^* \leftarrow \text{IBE.Enc}(PK, id^*, m)$.
4. \mathcal{A}_2 may continue querying oracle $\text{IBE.Ext}(SK, \cdot)$, with the restriction that it can not submit the challenge identity id^* . Finally, \mathcal{A} outputs a plaintext m' .

We say that \mathcal{A} succeeds if $m' = m$, and denote the adversary's advantage as $\Pr_{\mathcal{A}, \text{IBE}}[\text{OW}]$.

2.2 Indistinguishability Against Chosen-Ciphertext Attacks for IBE

In the following we recall the chosen-ciphertext indistinguishability security notions for IBE considered in the literature.

Definition 3. Let $\Pi = (\text{IBE.Gen}, \text{IBE.Ext}, \text{IBE.Enc}, \text{IBE.Dec})$ be an IBE scheme, and let $\mathcal{A} = (\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2)$ be any 3-tuple of PPT oracle algorithms. We say Π is IND-sID-CCA secure if for any 3-tuple of PPT oracle algorithms \mathcal{A} and any polynomially-bounded function $\ell(\cdot)$, the advantage in the following game is negligible in the security parameter 1^k :

1. $\mathcal{A}_0(1^k, \ell(k))$ outputs a target identity id^* .
2. $\text{IBE.Gen}(1^k)$ outputs (PK, SK) . The adversary is given PK .
3. FIND PHASE: The adversary \mathcal{A}_1 may ask polynomially-many queries to an oracle $\text{IBE.Ext}(SK, \cdot)$, except that it can not ask for the secret key related to identity id^* . The adversary \mathcal{A}_1 may also ask a decryption oracle $\text{IBE.Dec}(SK, \cdot, \cdot)$ for pairs identity-ciphertext (id, c) of its choice.
4. At some point, \mathcal{A}_1 outputs two equal length messages m_0, m_1 . A bit $b \leftarrow \{0, 1\}$ is chosen at random and the adversary is given a challenge ciphertext $c^* \leftarrow \text{IBE.Enc}(PK, id^*, m_b)$.
5. GUESS PHASE: \mathcal{A}_2 may continue querying oracle $\text{IBE.Ext}(SK, \cdot)$, with the restriction that it can not submit the challenge identity id^* . Again, \mathcal{A}_2 may also ask a decryption oracle $\text{IBE.Dec}(SK, \cdot, \cdot)$ for pairs identity-ciphertext (id, c) of its choice, with the restriction $(id, c) \neq (id^*, c^*)$. Finally, \mathcal{A}_2 outputs a guess b' .

We say that \mathcal{A} succeeds if $b' = b$, and define its advantage by $\Pr_{\mathcal{A}, \text{IBE}}[\text{sID}] = |\Pr[b' = b] - 1/2|$. Chosen-ciphertext security against full-identity attacks (IND-ID-CCA) is defined similarly, with the exception that the target identity is chosen by \mathcal{A}_1 at the end of the find phase.

3 Definitions for Tag-Based Encryption

Definition 4. A tag-based encryption scheme handling tags of length ℓ (where ℓ is a polynomially-bounded function) is specified by three PPT algorithms:

Setup. TBE.Gen takes a security parameter 1^k and returns a public key PK and secret key SK . The public key PK include the security parameter 1^k and $\ell(k)$; as well as the description of sets \mathcal{M}, \mathcal{C} , which denote the set of messages and ciphertexts respectively.

Encrypt. TBE.Enc takes as inputs PK , a tag $t \in \{0, 1\}^\ell$ and $m \in \mathcal{M}$. It returns a ciphertext $c \in \mathcal{C}$.

Decrypt. TBE.Dec takes as inputs the secret key SK and $c \in \mathcal{C}$, and it returns $m \in \mathcal{M}$ or reject when c is not a legitimate ciphertext. For the sake of consistency, these algorithms must satisfy $\text{TBE.Dec}(PK, t, c) = m$ for all $t \in \{0, 1\}^\ell, m \in \mathcal{M}$, where $c = \text{TBE.Enc}(PK, t, m)$.

3.1 One-Wayness for TBE

In the following, the notion of one-wayness against adaptive-tag chosen-plaintext attacks (referred to as OW-TBE-CPA in the following definition) is given.

Definition 5. Let $\mathcal{E} = (\text{TBE.Gen}, \text{TBE.Enc}, \text{TBE.Dec})$ be a TBE scheme, and let $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ be any 2-tuple of PPT oracle algorithms. We say \mathcal{E} is OW-TBE-CPA secure if for any 2-tuple of PPT oracle algorithms \mathcal{A} and any polynomially-bounded function $\ell(\cdot)$, the advantage in the following game is negligible in the security parameter 1^k :

1. $\text{TBE.Gen}(1^k)$ outputs (PK, SK) . The adversary is given PK .
2. The adversary \mathcal{A}_1 outputs a challenge tag t^* .
3. A message m is randomly chosen in \mathcal{M} and the adversary is given a challenge ciphertext $c^* \leftarrow \text{TBE.Enc}(PK, t^*, m)$.
4. Finally, \mathcal{A}_2 outputs a plaintext m' .

We say that \mathcal{A} succeeds if $m' = m$, and denote the adversary's advantage as $\Pr_{\mathcal{A}, \text{TBE}}[\text{OW}]$.

3.2 Indistinguishability Against Chosen Ciphertext Attacks for TBE

In the following we recall the chosen-ciphertext indistinguishability security notions for TBE considered in the literature.

Definition 6. Let $\mathcal{E} = (\text{TBE.Gen}, \text{TBE.Enc}, \text{TBE.Dec})$ be a TBE scheme, and let $\mathcal{A} = (\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2)$ be any 3-tuple of PPT oracle algorithms. We say \mathcal{E} is IND-sTBE-CCA secure if for any 3-tuple of PPT oracle algorithms \mathcal{A} and any polynomially-bounded function $\ell(\cdot)$, the advantage in the following game is negligible in the security parameter 1^k :

1. $\mathcal{A}_0(1^k, \ell(k))$ outputs a target tag t^* .
2. $\text{TBE.Gen}(1^k)$ outputs (PK, SK) . The adversary is given PK .
3. FIND PHASE: The adversary \mathcal{A}_1 may ask polynomially-many queries to a decryption oracle $\text{TBE.Dec}(SK, t, c)$ for pairs tag-ciphertext (t, c) of its choice, with the restriction $t \neq t^*$.
4. At some point, \mathcal{A}_1 outputs two equal length messages m_0, m_1 . A bit $b \leftarrow \{0, 1\}$ is chosen at random and the adversary is given a challenge ciphertext $c^* \leftarrow \text{TBE.Enc}(PK, t^*, m_b)$.
5. GUESS PHASE: \mathcal{A}_2 may continue asking the decryption oracle for pairs tag-ciphertext (t, c) of its choice, with the restriction $t \neq t^*$. Finally, \mathcal{A}_2 outputs a guess b' .

We say that \mathcal{A} succeeds if $b' = b$, and define its advantage by $\Pr_{\mathcal{A}, \text{TBE}}[\text{sTBE}] = |\Pr[b' = b] - 1/2|$. Chosen-ciphertext security against adaptive-tag attacks (IND-TBE-CCA) is defined similarly, with the exception that \mathcal{A}_1 is not placed any restriction, and the target identity is chosen by \mathcal{A}_1 at the end of the find phase.

4 A Separation Between Selective-Identity and Full-Identity Security Notions

In this section we present a separation between the selective-identity and full-identity security notions for IBE. More concretely, we show that IND-sID-CCA security *fails to imply* OW-ID-CPA security.

Theorem 1. IND-ID-CCA security implies OW-ID-CPA security.

Proof: The proof of this theorem is straightforward and therefore the details are omitted here. □

Theorem 2. IND-sID-CCA security does not imply OW-ID-CPA security.

Proof: Assume that there exists an IBE scheme $\Pi = (\text{IBE.Gen}, \text{IBE.Ext}, \text{IBE.Enc}, \text{IBE.Dec})$ which is IND-sID-CCA secure (otherwise the claim is trivially true). We construct another IBE scheme $\Pi' = (\text{IBE.Gen}', \text{IBE.Ext}', \text{IBE.Enc}', \text{IBE.Dec}')$ which is IND-sID-CCA secure but not OW-ID-CPA secure, whose existence proves the theorem. The scheme Π' is defined as follows:

IBE.Gen (1^k)	IBE.Ext (SK, id)	IBE.Enc (PK, id, m)	IBE.Dec (d_{id}, c)
$(PK, SK) \leftarrow \text{IBE.Gen}(1^k);$ $id^+ \leftarrow \{0, 1\}^{\ell};$ $d^+ \leftarrow \text{IBE.Ext}(SK, id^+);$ $PK \leftarrow (PK, id^+, d^+);$ return (PK, SK)	return $\text{IBE.Ext}(SK, id);$	return $\text{IBE.Enc}(PK, id, m)$	return $\text{IBE.Dec}(d_{id}, c)$

where $\ell' = \ell'(k)$ is a polynomially-bounded function. It is trivial to check that Π' qualifies as an IBE scheme if Π does. Π' is not OW-ID-CPA due to the following successful adversary $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$:

Algorithm $\mathcal{A}_1(PK)$ return id^+	Algorithm $\mathcal{A}_2(c)$ return $\text{IBE.Dec}(d^+, c)$
---	--

A simple calculation shows that the OW-ID-CPA advantage of the adversary $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ is 1. The basic idea is that \mathcal{A}_1 knows the decryption key related to id^+ once it gets PK' . Then it sets $id^* := id^+$, it qualifies as a OW-ID-CCA adversary (\mathcal{A}_1 did not query id^+ to its oracle $\text{IBE.Ext}(SK, \cdot)$) and finally it can decrypt any ciphertext related to the challenge identity, thus effectively breaking the one-wayness of Π' . It remains to show that the new scheme Π' is secure in the sense of IND-sID-CCA.

We argue by contradiction: let $\mathcal{C} = (\mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_2)$ be a 3-tuple algorithm breaking the IND-sID-CCA security of the IBE scheme $\Pi' = (\text{IBE.Gen}', \text{IBE.Ext}', \text{IBE.Enc}', \text{IBE.Dec}')$. Then there exists a 3-tuple algorithm $\mathcal{B} = (\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_2)$ whose success probability against the IND-sID-CCA security of the original IBE scheme Π is the same that \mathcal{C} has with respect to Π' . This implies a contradiction with the claim that Π is secure in the sense of IND-sID-CCA. The algorithm \mathcal{B} uses \mathcal{C} as a subroutine and is defined as follows:

Algorithm $\mathcal{B}_0(1^k, \ell)$ $id \leftarrow \mathcal{C}_0(1^k, \ell);$ return id	Algorithm $\mathcal{B}_1(PK, id)$ $id \neq id^+ \leftarrow \{0, 1\}^{\ell};$ $d^+ \leftarrow \text{IBE.Ext}(SK, id^+);$ $PK \leftarrow (PK, id^+, d^+);$ $(m_0, m_1) \leftarrow \mathcal{C}_1(PK, id);$ return (m_0, m_1)	Algorithm $\mathcal{B}_2(id, c)$ $v \leftarrow \mathcal{C}_2(id, c);$ return v
---	---	---

Notice that \mathcal{B} can easily simulate the oracles $\text{IBE.Ext}'(SK, \cdot)$ and $\text{IBE.Dec}'(SK, \cdot, \cdot)$ for \mathcal{C} by using its own oracles. By construction, the IND-sID-CCA advantages of \mathcal{C} and \mathcal{B} against its respective schemes are the same. \square

Theorems 1 and 2 imply that the strongest selective-identity security level (chosen-ciphertext security) is strictly weaker than the weakest full-identity security level (one-wayness).

5 A Separation Between Selective-Tag and Adaptive-Tag Security Notions

In this section we present a separation between the selective-tag and adaptive-tag security notions for TBE. That is, similarly to the previous section, we show that IND-sTBE-CCA security *fails to imply* OW-TBE-CPA security.

However, it does not seem easy to apply the technique used in Theorem 2 to the TBE case, the reason being that in a TBE scheme the only secret information we can publish in the public key PK is the secret key SK , which would result in a completely insecure scheme. For this reason, a different technique seems to be needed.

Theorem 3. IND-TBE-CCA security *implies* OW-TBE-CPA security.

Proof: The proof of this theorem is straightforward and therefore the details are omitted here. \square

Theorem 4. IND-sTBE-CCA security *does not imply* OW-TBE-CPA security.

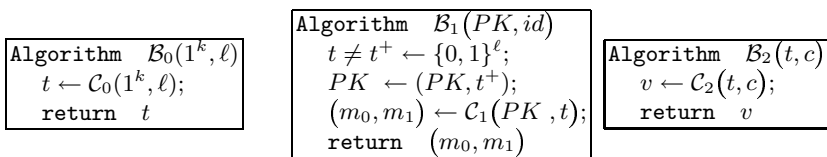
Proof: Assume that there exists a TBE scheme $\mathcal{E} = (\text{TBE.Gen}, \text{TBE.Enc}, \text{TBE.Dec})$ which is IND-sTBE-CCA secure. We construct another TBE scheme $\mathcal{E}' = (\text{TBE.Gen}', \text{TBE.Enc}', \text{TBE.Dec}')$ which is IND-sTBE-CCA secure but not OW-TBE-CPA secure, whose existence proves the theorem. The scheme \mathcal{E}' is defined as follows:

$\text{TBE.Gen}(1^k)$ $(PK, SK) \leftarrow \text{TBE.Gen}(1^k);$ $t^+ \leftarrow \{0, 1\}^\ell;$ $PK \leftarrow (PK, t^+);$ return (PK, SK)	$\text{TBE.Enc}(PK, t, m)$ Parse PK as (PK, t^+) $y \leftarrow \text{IBE.Enc}(PK, t, m)$ if $t = t^+$ then $c \leftarrow (y, m)$ else $r \leftarrow \mathcal{M}$ $c \leftarrow (y, r)$ return c	$\text{TBE.Dec}(SK, t, c)$ Parse c as (y, z) return $\text{TBE.Dec}(SK, t, y)$
--	---	--

where $\ell' = \ell'(k)$ is a polynomially-bounded function. It is trivial to check that \mathcal{E}' qualifies as an IBE scheme if \mathcal{E} does. In particular, the new ciphertext space \mathcal{C}' for \mathcal{E}' is $\mathcal{C}' = \mathcal{C} \times \mathcal{M}$. The scheme \mathcal{E}' is not OW-TBE-CPA due to the following successful adversary $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$. Once \mathcal{A}_1 gets PK' it outputs t^+ as the

target tag. For any $m \in \mathcal{M}$ it turns out that $(y, m) \leftarrow \text{TBE.Enc}'(PK', t^+, m)$ by definition of \mathcal{E}' . Finally, \mathcal{A}_2 outputs m on input (y, m) . A simple calculation shows that the OW-TBE-CPA advantage of the adversary $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ is 1.

It remains to show that the new scheme \mathcal{E}' is secure in the sense of IND-sTBE-CCA. Intuitively, an IND-sTBE-CCA adversary against \mathcal{E}' must guess ahead of time which will be the “doomy” tag t^+ . But this is infeasible, since this tag is chosen uniformly at random by the challenger. The details follow. We argue by contradiction: let $\mathcal{C} = (\mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_2)$ be a 3-tuple algorithm breaking the IND-sTBE-CCA security of the scheme \mathcal{E}' . Then there exists a 3-tuple algorithm $\mathcal{B} = (\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_2)$ whose success probability against the IND-sTBE-CCA security of the original IBE scheme \mathcal{E} is the same that \mathcal{C} has with respect to \mathcal{E}' .



Notice that \mathcal{B} can easily simulate the oracle $\text{TBE.Dec}'(SK, \cdot, \cdot)$ to \mathcal{C} by using its own oracle $\text{TBE.Dec}(SK, \cdot, \cdot)$. In addition, the challenge ciphertext c^* for \mathcal{C} is obtained by taking $r \leftarrow \mathcal{M}$ and setting $c^* = (c_*, r)$, where c_* is \mathcal{B} 's challenge ciphertext. By construction, the IND-sTBE-CCA advantages of \mathcal{C} and \mathcal{B} against its respective schemes are equal. □

Theorems 3 and 4 imply that the strongest selective-tag security level is strictly weaker than the weakest adaptive-tag security level.

6 Conclusions

In this note we have shown that in identity-based encryption the strongest selective-identity security level fails to imply the weakest full-identity security level. Therefore, selective-identity security turns out to be a strictly weaker security requirement than full-identity security. Previously, selective-identity security was regarded only as a *slightly weaker* security requirement than adaptive-identity security [BB04]. We have also presented a separation result for the related cryptographic primitive of tag-based encryption.

Acknowledgements. We are thankful to Eike Kiltz for providing us with an improved separation result for IBE and for stimulating discussions.

References

[BB04] D. Boneh and X. Boyen. Efficient selective-ID secure identity-based encryption without Random Oracles. In *EUROCRYPT 2004*, vol. 3027 of *LNCS*, pp. 223–238, 2004.

- [BF03] D. Boneh and M. Franklin. Identity-Based encryption from the Weil pairing. *SIAM Journal of Computing*, 32(3):586–615, 2003. This is the full version of an extended abstract of the same title presented at *Crypto'01*.
- [BR93] M. Bellare and P. Rogaway. Random oracles are practical: A paradigm for designing efficient protocols. In *Proceedings of the 1st ACM CCS*, pp. 62–73. ACM Press, 1993.
- [CHK04] R. Canetti, S. Halevi and J. Katz. Chosen-ciphertext security from identity-based encryption. In *EUROCRYPT 2004*, vol. 3027 of *LNCS*, pp. 207–222, 2004.
- [GM84] S. Golwasser and S. Micali. Probabilistic encryption. *Journal of Computer and System Sciences*, 28:270–299, 1984.
- [Kil06] E. Kiltz. Chosen-ciphertext security from tag-based encryption. In *Theory of Cryptography Conference 2006*, vol. 3876 of *LNCS*, pp. 581–600, 2006.
- [MRY04] P.D. MacKenzie, M.K. Reiter and K. Yang. Alternatives to non-malleability: Definitions, constructions, and applications (extended abstract). In *Theory of Cryptography Conference 2004*, vol. 2951 of *LNCS*, pp. 171–190, 2004.
- [NY90] M. Naor and M. Yung. Public-key cryptosystems provably secure against chosen ciphertext attack. In *Proc. of the Twenty-Second Annual ACM Symposium on Theory of Computing*, pp. 427–437. ACM, 1990.
- [RS92] C. Rackoff and D.R. Simon. Non-interactive zero-knowledge proof of knowledge and chosen ciphertext attack. In *CRYPTO 1991*, vol. 576 of *LNCS*, pp. 433–444, 1992.
- [Sha85] A. Shamir. Identity-based cryptosystems and signature schemes. In *CRYPTO 1984*, vol. 196 of *LNCS*, pp. 47–53, 1985.

Traceable Signature: Better Efficiency and Beyond*

He Ge and Stephen R. Tate

Dept. of Computer Science and Engineering, University of North Texas
ge@unt.edu
srt@cs.unt.edu

Abstract. In recent years one of the most active research areas in applied cryptography is the study of techniques for creating a group signature, a cryptographic primitive that can be used to implement anonymous authentication. Some variants of group signature, such as traceable signature, and authentication with variable anonymity in a trusted computing platform, have also been proposed. In this paper we propose a traceable signature scheme with variable anonymity. Our scheme supports two important properties for a practical anonymous authentication system, i.e., corrupted group member detection and fair tracing, which have unfortunately been neglected in most group signature schemes in the literature. We prove the new scheme is secure in the random oracle model, under the strong RSA assumption and the decisional Diffie-Hellman assumption.

Keywords: Group signature, Traceable Signature, Anonymous Authentication, Variable Anonymity, Cryptographic Protocol.

1 Introduction

In this paper, we present new techniques for performing anonymous authentication, in which authenticated users receive credentials from a designated group manager, and in later interactions a user can prove possession of such a credential in a privacy-preserving manner. Anonymous authentication has been one of the most active research areas in applied cryptography in recent years.

The most heavily studied type of anonymous authentication system is the “group signature scheme,” which provides a well-defined set of services and security guarantees that we describe in more detail below.¹ However, several authors have identified various desirable properties not provided by the group signature definition, and have introduced variants of this basic scheme including work on “anonymous credential systems” [6], “traceable signatures” [14], and a system designed for trusted computing platforms called “direct anonymous attestation” [4]. Our contribution in this paper is to show how the well-known group signature scheme of Ateniese *et al.* [1], which we call the ACJT scheme, can be modified to a traceable signature so that it supports a particularly useful extension from the work on direct anonymous attestation that allows a

* This research is supported in part by NSF award 0208640.

¹ An extensive bibliography of group signature literature can be found at <http://www.i2r.star.edu.sg/icsd/staff/guilin/bible/group-sign.htm>

prover and verifier to agree on a variable degree of signature linkability. Our modifications to the ACJT scheme replace operations with modified formulas that have the same computational complexity, so our system preserves the efficiency of the ACJT scheme while providing a unique set of features which is useful in many situations.

1.1 Background

Group signature is a privacy-preserving signature scheme introduced by Chaum and Heyst in 1991 [10]. In such a scheme, a group member can sign a message on behalf of the group without revealing his identity. Only the group manager or the specified open authority can open a signature and find its originator. Signatures made by the same user cannot be identified as from the same source, i.e., “linked”. Recently, group signature has attracted considerable attention, and many schemes have been proposed in the literature (e.g., [8, 1, 6, 7, 5]). Creating an anonymous authentication scheme from a group signature is simple: the group is simply the set of authorized users, and authentication is performed by a group member placing a group signature on a challenge (nonce) sent by the service requiring authentication. From the properties of group signatures, all the service or an attacker can learn is that the signature was made by a valid group member (i.e., an authorized user).

However, group signature does not provide certain important features for a more hostile or realistic environment where group members could be malicious or compromised. In such settings, an efficient mechanism should be available to reveal all the malicious behaviors of corrupted members. In group signature, identification of signatures from corrupted members has to be done by opening all signatures. This is either inefficient (centralized operation by the group manager), or unfair (unnecessarily identifying all innocent group members’ signatures). To overcome this shortcoming, Kiayias *et al.* proposed a variant of group signature, called traceable signature [14]. They define “traceability” as the ability to reveal all the signatures signed by a group member without requiring the open authority to open them. Tracing can be done by “trace agents” distributively and efficiently. They also introduced the concept of “self-traceability”, or “claiming”. That is, a group member himself can stand out, claiming a signature signed by himself without compromising his other signatures and secrets. The subtlety lies in that a group member should be able to do this without keeping all one time random values in his signatures. In group signature, a group member may also be able to claim his signatures, but he has to keep all his transaction transcripts including some random values, making “claiming” highly impractical and a security risk.

The Trusted Computing Group [15] has recently proposed an architecture called the “trusted computing platform” to enhance computer security. A trusted computing platform is a computing device integrated with a cryptographic chip called the trusted platform module (TPM). The TPM is designed and manufactured so that all other remote parties can trust cryptographic computing results from this TPM. To protect the privacy of a TPM owner, an anonymous authentication technique, called Direct Anonymous Attestation (DAA), has been deployed in recent versions of the trusted computing platform. DAA can be seen as a group signature scheme without openability. DAA introduces the notion of “variable anonymity,” which is conditionally linkable anonymous authentication: the same TPM will produce linkable signatures for a certain period of

time. The period of time during which signatures can be linked can be determined by the parties involved and can vary from an infinitesimally short period (leading to completely unlinkable signatures) to an infinite period (leading to completely linkable signatures). Signatures made by the same user in different periods of time or to different servers cannot be linked. By setting the linkability period to a moderately short time period (a day to a week) a server can potentially detect if a key has been compromised and is being used by many different users, while still offering some amount of unlinkability.

1.2 Our Results

In the previous section we briefly introduced some of the available techniques for anonymous authentication. Numerous constructions with different features have been proposed to accommodate different properties. This raised the question which we address in this paper: Can we devise a construction which combines the features from different authentication primitives? More specifically, can we have a traceable signature scheme which also supports variable anonymity? So far as we know, no such scheme has been proposed to work in this manner, probably because variable anonymity is a recently identified feature in anonymous authentication.

We consider the combination of traceability and variable anonymity to be particularly important for anonymous authentication. Variable anonymity is the only way key sharing violations can be detected, while traceability is the efficient and fair way to reveal all malicious behaviors. More specifically, while the standard group signature scheme can use the open authority to identify a user that performs malicious actions, consider what happens when one authorized user shares his authentication credential with a set of co-conspirators. For example, a large set of users could share a single subscription to some pay web site. Since all authentications are completely unlinkable in a group signature scheme, it would be impossible to determine whether 1000 requests coming in during a day are from 1000 different valid users or from 1000 people sharing a single valid credential. Introducing linkability for a limited time period is the only way to detect this, and if an unusually high number of requests using the same credential come in from different IP addresses during the same day, then this could be flagged as potentially malicious behavior. After that, the open authority can open the signatures to determine the real owner of this credential, and the tracing trapdoor associated with this credential is further revealed to trace agents by the group manager. Then the trace agents reveal all the behaviors associated with the trapdoor for further investigation. At the same time, a tracing trapdoor may be published on the revocation list for verifiers to identify future requests by this member. In our opinion, to build up a realistic anonymous authentication system, the combination of traceability and variable anonymity is a must.

In this paper, we present our construction for traceable signature that supports variable anonymity. Our construction is built up from the well-known ACJT group signature [1]. The traceable signature due to Kiayias *et al.*, which we refer to as the KTY scheme in this paper, is also built up from the ACJT scheme. However, our construction improves on the KTY scheme in three aspects. First, we adopt the same group membership certificate as in the ACJT scheme. The KTY scheme changes the group certificate in the ACJT scheme to integrate the tracing trapdoor. We show this change is unnecessary by identifying that tracing trapdoors in fact are already available in the ACJT scheme. Second, our

tracing mechanism is more efficient than the KTY scheme. Our scheme uses a hash function to create generators while the KTY scheme uses expensive exponentiation computation. Finally, our scheme supports variable anonymity while the KTY scheme does not. Thus, our scheme is more efficient and flexible than the KTY scheme.

The rest of this paper is organized as follows. The next section introduces a concrete model for our signature scheme. Section 3 reviews some definitions, cryptographic assumptions, and building blocks of our proposed scheme. Section 4 presents the proposed scheme. Security properties are considered in Section 5. Finally, we summarize and give conclusions in section 6.

2 The Model

This section introduces the model for traceable signature [14], which is a variant of the group signature model (e.g. [1]). Both of these two models include operations for Setup, Join, Sign, Verify, and Open. The traceable signature model has additional operations for traceability: Reveal, Trace, Claim (Self-trace) and Claim-Verify.

Definition 1. *A traceable signature is a digital signature scheme with four types of participants: Group Manager, Group Members, Open Authorities, and Trace Agents. It consists of the following procedures:*

- **Setup:** *For a given security parameter σ , the group manager produces system-wide public parameters and a group manager master key for group membership certificate generation.*
- **Join:** *An interactive protocol between a user and the group manager. The user obtains a group membership certificate to become a group member. The public certificate and the user's identity information are stored by the group manager in a database for future use.*
- **Sign:** *Using its group membership certificate and private key, a group member creates a group signature for a message.*
- **Verify:** *A signature is verified to make sure it originates from a legitimate group member without the knowledge of which particular one.*
- **Open:** *Given a valid signature, an open authority discloses the underlying group membership certificate.*
- **Reveal:** *The group manager outputs the tracing trapdoor associated with a group membership certificate.*
- **Trace:** *Trace agents check whether a signature is associated with a tracing trapdoor.*
- **Claim (Self-trace):** *A group member creates a proof that he created a particular signature.*
- **Claim-Verify:** *A party verifies the correctness of the claiming transcript.*

Similar to group signatures, a traceable signature scheme should satisfy the following properties:

- **Correctness:** Any valid signature can be correctly verified by the Verify protocol and a valid claiming proof can be correctly verified.

- **Forgery-Resistance:** A valid group membership certificate can only be created by a user and the group manager through Join protocol.
- **Anonymity:** It is infeasible to identify the real signer of a signature except by the open authority or if the signature has been claimed.
- **Unlinkability:** It is infeasible to link two different signatures of the same group member.
- **Non-framing:** No one (including the group manager) can sign a message in such a way that it appears to come from another user if it is opened.
- **Traceability:** Given a tracing trapdoor, trace agents can reveal all signatures associated with the trapdoor. A group member can claim (self-trace) his signatures.

3 Definitions and Preliminaries

This section reviews some definitions, widely accepted complexity assumptions that we will use in this paper, and building blocks for our construction.

Definition 2 (Special RSA Modulus). *An RSA modulus $n = pq$ is called special if $p = 2p' + 1$ and $q = 2q' + 1$ where p' and q' also are prime numbers.*

Definition 3 (Quadratic Residue Group QR_n). *Let Z_n^* be the multiplicative group modulo n , which contains all positive integers less than n and relatively prime to n . An element $x \in Z_n^*$ is called a quadratic residue if there exists an $a \in Z_n^*$ such that $a^2 \equiv x \pmod{n}$. The set of all quadratic residues of Z_n^* forms a cyclic subgroup of Z_n^* , which we denote by QR_n . If n is the product of two distinct primes, then $|QR_n| = \frac{1}{4}|Z_n^*|$.*

The security of our techniques relies on the following security assumptions which are widely accepted in the cryptography literature (see, for example, [2, 13, 8, 1]).

Assumption 1 (Strong RSA Assumption). *Let n be an RSA modulus. The Flexible RSA Problem is the problem of taking a random element $u \in Z_n^*$ and finding a pair (v, e) such that $e > 1$ and $v^e = u \pmod{n}$. The Strong RSA Assumption says that no probabilistic polynomial time algorithm can solve the flexible RSA problem with non-negligible probability.*

Assumption 2. (Decisional Diffie-Hellman Assumption for QR_n) *Let n be a special RSA modulus, and let g be a generator of QR_n . For the two distributions (g, g^x, g^y, g^{xy}) , (g, g^x, g^y, g^z) , $x, y, z \in_R Z_n$, there is no probabilistic polynomial-time algorithm that distinguishes them with non-negligible probability.*

The building blocks of our technique are *statistical honest-verifier zero knowledge proofs of knowledge* related to discrete logarithms over QR_n [9, 8]. They may include protocols for problems such as the knowledge of the discrete logarithm, the knowledge of equality of two discrete logarithms, the knowledge of the discrete logarithm that lies in certain interval, etc. We introduce one of them here. Readers may refer to the original papers for more details.

Protocol 1. Let n be a special RSA modulus, QR_n be the quadratic residue group modulo n , and g be a generator of QR_n . ϵ, l, l_c are security parameters that are all greater than 1. X is a constant number. A prover Alice knows x , the discrete logarithm of T_1 , and $x \in [X - 2^l, X + 2^l]$. Alice demonstrates her knowledge of $x \in [X - 2^{\epsilon(l+l_c)}, X + 2^{\epsilon(l+l_c)}]$ as follows.

1. Alice picks a random $t \in \pm\{0, 1\}^{\alpha(l+l_c)}$ and computes $T_2 = g^t \pmod{n}$. Alice sends (T_1, T_2) to a verifier Bob.
2. Bob picks a random $c \in \{0, 1\}^{l_c}$ and sends it to Alice.
3. Alice computes $w = t - c(x - X)$, and $w \in \pm\{0, 1\}^{\alpha(l+l_c)+1}$. Alice sends w to Bob.
4. Bob checks $w \in \pm\{0, 1\}^{\alpha(l+l_c)+1}$ and

$$g^{w-cX}T_1^c \stackrel{?}{=} T_2 \pmod{n}.$$

If the equation holds, Alice proves knowledge of the discrete logarithm of T_1 lies in the range $[X - 2^{\epsilon(l+l_c)}, X + 2^{\epsilon(l+l_c)}]$.

Remark 1. It should be emphasized that while Alice knows a secret x in $[X - 2^l, X + 2^l]$, the protocol only guarantees that x lies in the extended range $[X - 2^{\epsilon(l+l_c)}, X + 2^{\epsilon(l+l_c)}]$.

Remark 2. Using the Fiat-Shamir heuristic [12], the protocol can be turned into a non-interactive “signature of knowledge” scheme, which is secure in the random oracle model [3]. We will introduce the proposed scheme in the manner of “signature of knowledge” in next section.

4 Traceable Signature

Our construction is built upon the ACJT group signature scheme. We adopt the same system parameters, group certificates, and Join protocol. The Sign and Verify protocols have been changed to support traceability and variable anonymity. In the following presentation, we use the same notation as in the original paper to make it easier for readers to see how we convert the ACJT scheme into a traceable signature scheme.

4.1 The System Parameters

The following system parameters are set up when the system is initialized and the group manager key is generated.

- A special RSA modulus $n = pq$, $p = 2p' + 1$, $q = 2q' + 1$, with p, p', q, q' all prime
- Random elements $a, a_0, g \in QR_n$ of order $p'q'$, i.e., these numbers are generators of QR_n
- Security parameters used in protocols: $\epsilon > 1, k, l_p$
- Length parameters $\lambda_1, \lambda_2, \gamma_1, \gamma_2$. $\lambda_1 > \epsilon(\lambda_2 + k) + 2$, $\lambda_2 > 4l_p$, $\gamma_1 > \epsilon(\gamma_2 + k) + 2$, and $\gamma_2 > \lambda_1 + 2$
- Integer ranges $\Lambda =]2^{\lambda_1} - 2^{\lambda_2}, 2^{\lambda_1} + 2^{\lambda_2}[$ and $\Gamma =]2^{\gamma_1} - 2^{\gamma_2}, 2^{\gamma_1} + 2^{\gamma_2}[$
- Three strong collision-resistant hash functions: $\mathcal{H}_1, \mathcal{H}_2 : \{0, 1\}^* \rightarrow Z_n^*$, and $\mathcal{H}_3 : \{0, 1\}^* \rightarrow \{0, 1\}^k$

- A message to be signed: $m \in \{0, 1\}^*$
- The public parameters are (n, a, a_0, g) .
- The secret parameters for the group manager are (p', q') .

The open authority creates his ElGamal public keypair [11], i.e., private key x and public key y such that $y = g^x \pmod n$.

4.2 Variable Anonymity Parameter

To achieve variable anonymity, each signature will belong to a “linkability class” that is identified using a “linkability class identifier,” or LCID. All signatures made by the same group member with the same LCID are linkable, and in an interactive authentication protocol the LCID can be negotiated and determined by the two parties. For example, to link authentications to a single server over a single day, the LCID could simply be the server name concatenated with the date. If the same LCID is always used with a particular server (e.g., the server name), then the result is a pseudo-anonymity system. If complete anonymity is desired, the signer can simply pick a random LCID (which is possible if the server isn’t concerned with linkability and allows arbitrary LCIDs).

4.3 Join Protocol

The same Join protocol is adopted as in the original scheme. A group membership certificate is in the form of $A_i = (a^{x_i} a_0)^{1/e_i} \pmod n$ where $x_i \in \Lambda$ is the secret of the group member, and $e_i \in_R \Gamma$ is a random prime number that is known to both the group member and group manager.²

In our scheme, e_i is treated as tracing trapdoor, and kept secret by the group member and group manager. When an open authority reveals A_i for a signature, the group manager sends the corresponding e_i to the trace agents in order to trace all signatures associated with e_i .

x_i is treated as self-tracing trapdoor, which is used by a group member to claim his signatures. Since x_i is the secret of group member, only group member himself have the ability to claim his signatures.

4.4 Sign Protocol

In order to sign a message m , a group member does the following:

- Derive two generators i and j of QR_n by hashing the LCID of this signature.

$$i = (\mathcal{H}_1(LCID))^2 \pmod n, \quad j = (\mathcal{H}_2(LCID))^2 \pmod n.$$

In the random oracle model, with the hash functions modeled by random oracles, each distinct LCID results in i and j being random generators of QR_n with overwhelming probability.

² Kiayias *et al.* have showed the range of x_i, e_i can be much smaller without compromising the scheme’s security [14]. For simplicity, we still follow the definition in ACJT scheme.

- Generate a random value $w \in_R \{0, 1\}^{2l_p}$ and compute:

$$T_1 = A_i y^w \pmod{n}, T_2 = g^w \pmod{n}, T_3 = i^{e_i} \pmod{n}, T_4 = j^{x_i} \pmod{n}$$

- Randomly (uniformly) choose $r_1 \in_R \pm\{0, 1\}^{\epsilon(\gamma_2+k)}$, $r_2 \in_R \pm\{0, 1\}^{\epsilon(\lambda_2+k)}$, and $r_3 \in_R \pm\{0, 1\}^{\epsilon(\lambda_1+2l_p+k+1)}$, and compute
 - $d_1 = T_1^{r_1}/(a^{r_2} y^{r_3}) \pmod{n}$, $d_2 = T_2^{r_1}/g^{r_3} \pmod{n}$, $d_3 = i^{r_1} \pmod{n}$, $d_4 = j^{r_2} \pmod{n}$.
 - $c = \mathcal{H}_3(g||i||j||y||a_0||a||T_1||T_2||T_3||d_1||d_2||d_3||d_4||m)$;
 - $s_1 = r_1 - c(e_i - 2^{\gamma_1})$, $s_2 = r_2 - c(x_i - 2^{\lambda_1})$, $s_3 = r_3 - ce_i w$ (all in Z_n).
- Output the signature tuple $(LCID, c, s_1, s_2, s_3, T_1, T_2, T_3, T_4)$.

4.5 Verify Protocol

To verify a signature $(LCID, c, s_1, s_2, s_3, T_1, T_2, T_3, T_4)$, a verifier does the following.

- Compute the same generators i and j , and then

$$c' = \mathcal{H}_3(g||i||j||a_0||a||T_1||T_2||T_3||T_4||a_0^c T_1^{s_1 - c2^{\gamma_1}} / (a^{s_2 - c2^{\lambda_1}} y^{s_3}) || T_2^{s_1 - c2^{\gamma_1}} / g^{s_3} || i^{s_1 - c2^{\gamma_1}} T_3^c || j^{s_2 - c2^{\lambda_1}} T_4^c || m)$$

- Accept the signature if and only if $c = c'$ and $s_1 \in \pm\{0, 1\}^{\epsilon(\gamma_2+k)+1}$, $s_2 \in \pm\{0, 1\}^{\epsilon(\lambda_2+k)+1}$, $s_3 \in \pm\{0, 1\}^{\epsilon(\lambda_1+2l_p+k+1)+1}$.

4.6 Open and Reveal Protocol

For a valid signature, the open authority opens a signature to find its originator by ElGamal decryption:

$$A_i = T_1/T_2^x \pmod{n}.$$

For the non-framing property, the open authority must also issue a proof that it correctly revealed the group member, which can be done identically to the method used by the ACJT group signatures.

The opened certificate A_i is submitted to the group manager, and the group manager reveals the corresponding tracing trapdoor e_i to the trace agents.

4.7 Trace Protocol

To trace a group member, trace agents use e_i to reveal all the signatures by a group member by checking whether

$$i^{e_i} =? T_3 \pmod{n}.$$

To claim a signature, a group member proves its knowledge of discrete logarithm of T_4 with base j through Protocol 1.

5 Security Properties

Our scheme uses the same certificate as in the ACJT group signature. We have changed their Sign and Verify protocols. The security properties, such as, forgery-resistance, anonymity, non-framing, are unaffected by these changes. In this section, we only discuss the security properties affected by our change. Readers may refer to the original paper for other security arguments — the following theorem is representative, and further discussion is available in the full version of this paper.

Theorem 1 (Coalition-resistance). *Under the strong RSA assumption, a group certificate $[A_i = (a^{x_i} a_0)^{1/e_i} \pmod{n}, e_i]$ with $x \in \Lambda$ and $e_i \in \Gamma$ can be generated only by the group manager provided that the number K of certificates the group manager issues is polynomially bounded.*

Now, we address the security of Sign and Verify protocol, which is described as the following theorem.

Theorem 2. *Under the strong RSA assumption, and the decisional Diffie-Hellman assumption, the interactive protocol underlying the group signature scheme is a statistical zero-knowledge (honest-verifier) proof of knowledge of a membership certificate and a corresponding membership secret key.*

Proof. The proof for correctness is straightforward. A proof for the zero-knowledge property (simulator) following the same method in the KTY scheme (*Lemma 20*) appears in the full version of this paper. We only address the existence of a knowledge extractor, which is able to recover the group certificate when it has found two accepting tuples under the same commitment and different challenges from a verifier. Let $(T_1, T_2, T_3, d_1, d_2, d_3, c, s_1, s_2, s_3)$ and $(T_1, T_2, T_3, d_1, d_2, d_3, c', s'_1, s'_2, s'_3)$ be such tuples.

Since $d_4 \equiv j^{s_2 - c 2^{\lambda_1}} T_4^c \equiv j'^{s_2 - c} 2^{\lambda_1} T_4^c \pmod{n}$, we have

$$j^{(s_2 - s_2) + (c - c') 2^{\lambda_1}} \equiv T_4^{c - c'} \pmod{n}.$$

Under the strong RSA assumption, $c - c'$ has to divide $(s'_2 - s_2) + (c - c') 2^{\lambda_1}$. Therefore we have $\tau_1 = (s'_2 - s_2) / (c - c') + 2^{\lambda_1}$.

Since $d_3 \equiv i^{s_1 - c 2^{\gamma_1}} T_3^c \equiv i^{s_1 - c'} 2^{\gamma_1} T_3^c \pmod{n}$, we have

$$i^{(s_1 - s_1) + (c - c') 2^{\gamma_1}} \equiv T_3^{c - c'} \pmod{n}.$$

Likewise, under the strong RSA assumption, $c - c'$ has to divide $(s'_1 - s_1)$. We obtain $\tau_2 = (s'_1 - s_1) / (c - c') + 2^{\gamma_1}$.

Since $d_2 \equiv T_2^{s_1 - c 2^{\gamma_1}} / g^{s_3} \equiv T_2^{s_1 - c'} 2^{\gamma_1} / g^{s_3} \pmod{n}$, we have

$$T_2^{(s_1 - s_1) + (c - c') 2^{\gamma_1}} \equiv g^{s_3 - s_3} \pmod{n}.$$

Similarly, we have $\tau_3 = (s'_3 - s_3) / ((s'_1 - s_1) + (c - c') 2^{\gamma_1})$.

Since $d_1 \equiv a_0^c T_1^{s_1 - c 2^{\gamma_1}} / (a^{s_2 - c 2^{\lambda_1}} y^{s_3}) \equiv a_0^c T_1^{s_1 - c 2^{\gamma_1}} / (a^{s_2 - c 2^{\lambda_1}} y^{s_3}) \pmod{n}$, We have

$$a^{s_2 - s_2 + (c - c) 2^{\lambda_1}} a_0^{c - c} \equiv T_1^{s_1 - s_1 + (c - c) 2^{\gamma_1}} / y^{s_3 - s_3} \pmod{n}.$$

We further obtain

$$a^{(s_2 - s_2) / (c - c) + 2^{\lambda}} a_0 \equiv (T_1 / y^{(s_3 - s_3) / ((s_1 - s_1) + (c - c) 2^{\gamma_1})})^{(s_1 - s_1) / (c - c) + 2^{\gamma_1}} \pmod{n}.$$

Finally, let $A_i = T_1 / y^{\tau_3} \pmod{n}$, and then we obtain a valid certificate (A_i, τ_2, τ_1) such that $A_i^{\tau_2} = a^{\tau_1} a_0 \pmod{n}$, and τ_1, τ_2 lie in the valid range due to the length restriction on s_1, s_2, s_3 and c . Therefore we have demonstrated the existence of a knowledge extractor that can fully recover a valid group certificate. \square

Unlinkability follows the same argument in the ACJT group signature for T_1, T_2 . Since we define a new T_3, T_4 in our traceable signature, we need to show this change still keeps the unlinkability property (for different generators i and i'). Similar to the case in the ACJT group signature, the problem of linking two tuples $(i, T_3), (i', T'_3)$, is equivalent to deciding the equality of the discrete logarithms of T_3, T'_3 with base i, i' respectively. This is assumed to be infeasible under the decisional Diffie-Hellman assumption over QR_n . $(j, T_4), (j', T'_4)$ also follows the same argument. Therefore, we have the following result.

Theorem 3 (Unlinkability). *Under the decisional Diffie-Hellman assumption over QR_n and with \mathcal{H}_1 and \mathcal{H}_2 as random oracles, there exists no probabilistic polynomial-time algorithm that can make the linkability decision for any two arbitrary tuples $(i, T_3), (i', T'_3)$, or $(j, T_4), (j', T'_4)$ with non-negligible probability.*

6 Conclusion

We have presented a traceable signature scheme which is an enhancement of the ACJT group signature scheme [1] that supports variable anonymity. Our scheme is a more general solution to anonymous authentication, due to its support of traceability and variable anonymity. Traceability provides an efficient and fair mechanism to reveal and revoke corrupted group members, which is very important to a large, realistic anonymous authentication system. Variable anonymity can be adjusted to provide a wide range of linkability properties, from completely unlinkable signatures, to signatures linkable within a fixed time period, to completely linkable signatures (giving what is essentially a fixed pseudonym system). In practice, the amount of linkability would be determined by a risk analysis of the application, balancing the goal of protecting a user's privacy against a provider's goal of detecting inappropriate uses of keys. As our scheme supports the full range of linkability options, it provides the best available flexibility to users as well as providers. Finally, we have proved that our new signature scheme is secure under the strong RSA assumption and the Decisional Diffie-Hellman assumption over QR_n .

References

1. G. Ateniese, J. Camenisch, M. Joye, and G. Tsudik. A practical and provably secure coalition-resistant group signature scheme. In *Advances in Cryptology — Crypto*, pages 255–270, 2000.
2. N. Baric and B. Pfitzmann. Collision-free accumulators and fail-stop signature schemes without trees. In *Advances in Cryptology — Eurocrypt*, pages 480–494, 1997.
3. M. Bellare and P. Rogaway. Random oracles are practical: A paradigm for designing efficient protocols. In *ACM Conference on Computer and Communication Security*, pages 62–73, 1993.
4. E. Brickell, J. Camenisch, and L. Chen. Direct anonymous attestation. In *ACM Conference on Computer and Communications Security*, pages 132–145, 2004.
5. J. Camenisch and J. Groth. Group signatures: Better efficiency and new theoretical aspects. In *Security in Communication Networks (SCN 2004)*, LNCS 3352, pages 120–133, 2005.
6. J. Camenisch and A. Lysyanskaya. Dynamic accumulators and application to efficient revocation of anonymous credentials. In *Advances in Cryptology — Crypto'02*, LNCS 2442, pages 61–76, 2002.
7. J. Camenisch and A. Lysyanskaya. A signature scheme with efficient protocols. In *SCN'02*, LNCS 2576, pages 268–289, 2002.
8. J. Camenisch and M. Stadler. A group signature scheme with improved efficiency. In *Advances in Cryptology — ASIACRYPT'98*, LNCS 1514, pages 160–174, 1998.
9. A. Chan, Y. Frankel, and Y. Tsiounis. Easy come - easy go divisible cash. In K. Yyberg, editor, *Advances in Cryptology – Eurocrypt'98*, LNCS 1403, pages 561 – 574. Springer-Verlag, 1998.
10. D. Chaum and E. van Heyst. Group signature. In *Advances in Cryptology — Eurocrypt*, pages 390–407, 1992.
11. T. ElGamal. A public key cryptosystem and a signature scheme based on discrete logarithms. In *Advances in Cryptology — Crypto*, pages 10–18, 1984.
12. A. Fiat and A. Shamir. How to prove yourself: practical solutions to identification and signature problems. In *Advances in Cryptology — CRYPTO'86*, LNCS 263, pages 186–194. Springer-Verlag, 1987.
13. E. Fujisaki and T. Okamoto. Statistical zero knowledge protocols to prove modular polynomial relations. In *Advances in Cryptology — Crypto*, pages 16–30, 1997.
14. A. Kiayias, Y. Tsiounis, and M. Yung. Traceable signatures. In *Advances in Cryptology—Eurocrypt*, LNCS 3027, pages 571–589. Springer-Verlag, 2004.
15. TCG. <http://www.trustedcomputinggroup.org>.

On the TYS Signature Scheme

Marc Joye^{1,*} and Hung-Mei Lin²

¹ Gemplus, Security Technologies Department,
La Vigie, Avenue du Jujubier,
13705 La Ciotat Cedex, France

² Traverse des Jardins,
83640 Saint Zacharie, France

Abstract. This paper analyzes the security of a very efficient signature scheme proposed by C.H. Tan, X. Yi and C.K. Siew: the TYS signature scheme. We show that this scheme is universally forgeable; more specifically, we show that anyone is able to produce a valid TYS signature on a chosen message from an arbitrary valid message/signature pair. We also suggest modifications to the TYS signature scheme and relate the resulting scheme to the Camenisch-Lysyanskaya signature scheme.

Keywords: Cryptography, digital signature, standard model, TYS signature scheme, Camenisch-Lysyanskaya signature scheme.

1 Introduction

Designing secure yet efficient signature schemes is of central importance for cryptographic applications. The standard security notion for signature schemes is *existential unforgeability against adaptive chosen message attacks* (EUF-CMA) [12]. Informally, we require that an adversary getting access to a signing oracle and making adaptive signature queries on messages of her choice cannot produce a new valid signature. A scheme provably meeting this security notion is given in [12] with subsequent improvements in [7, 5]. More efficient, *stateless signature schemes* were later proposed. These include the Cramer-Shoup signature scheme [6] (revisited in [9]) and the Gennaro-Halevi-Rabin signature scheme [11], independently introduced in 1999.

More recently, C.H. Tan, X. Yi and C.K. Siew proposed a new signature scheme: the *TYS signature scheme* [16]. This scheme can be seen as an efficient variant of the Cramer-Shoup signature scheme, allowing faster signing and producing signatures roughly half the size.

Crosschecking the security proof offered in [16], we observed some inaccuracies. This illustrates once more that details are easily overlooked in security proofs [15] and this may have dramatic consequences.

We present in this paper an attack against the TYS signature scheme. We show how given a valid message/signature pair it is easy to produce a valid signature on a *chosen* message. Repairing the scheme does not seem possible.

* The work described in this paper has been supported in part by the European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT.

Instead, we suggest two simple modifications to the original scheme. Interestingly, the so-obtained scheme leads to an off-line/on-line variation [8, 14] of a previous scheme due to Camenisch and Lysyanskaya [4] (see also [17, 18]).

The rest of this paper is organized as follows. In the next section, we review the original TYS signature scheme. In Section 3, we show that the scheme is universally forgeable. Next, in Section 4, we suggest modifications to the original scheme. Finally, we conclude in Section 5.

2 Tan-Yi-Siew Signature Scheme

We review in this section the TYS signature scheme, as given in [16].

For a security parameter ℓ , let $H : \{0, 1\}^* \rightarrow \{0, 1\}^\ell$ be a collision-resistant hash function. The scheme consists of three algorithms: the key generation algorithm, the signing algorithm and the verification algorithm.

Key generation. Choose two random primes $p = 2p' + 1$ and $q = 2q' + 1$, where p' and q' are prime and $\log_2 p'q' > 2\ell + 1$, and let $N = pq$. Choose two quadratic residues g and x in \mathbb{Z}_N^* such that the order of g is $p'q'$. Finally, choose a random ℓ -bit integer z and compute $h = g^{-z} \pmod N$.
The public key is $\text{pk} = \{g, h, x, N\}$ and the private key is $\text{sk} = \{p, q, z\}$.

Signing. Let $m \in \{0, 1\}^*$ denote the message being signed. Randomly pick an ℓ -bit integer k and an ℓ -bit prime e . Next compute

$$y = (xg^{-k})^{1/e} \pmod N, \quad c = H(\text{pk}, y, m), \quad t = k + cz .$$

The signature on message m is $\sigma = (t, y, e)$.

Verification. Signature $\sigma = (t, y, e)$ on message m is accepted iff e is odd and

$$y^e g^t h^c \equiv x \pmod N$$

with $c' = H(\text{pk}, y, m)$.

Fig. 1. Original TYS signature scheme

The TYS signature scheme is very efficient and allows the use of coupons. The value of y can be precomputed off-line; only the pair (c, t) needs to be computed on-line.

3 Security Analysis

The authors of the TYS signature scheme conclude that their signature scheme is existentially unforgeable against chosen-message attacks (EUF-CMA) under the

strong RSA assumption. From the definition of a TYS signature, $\sigma = (t, y, e)$, we observe that $t = k + cz$ is not zero-knowledge if $k < cz$. We take advantage of this observation and present below a universal forgery.

Given a valid signature $\sigma = (t, y, e)$ on a message m , we compute

$$\hat{z} := \left\lfloor \frac{t}{c} \right\rfloor \quad \text{with } c = H(\text{pk}, y, m)$$

as an approximation for z . Since $t = k + cz$, we obviously have $z = \hat{z}$ when $k < c$. Otherwise, we have $z = \hat{z} - \delta$ for some integer $\delta > 0$. The correct value of δ can be guessed by checking whether $g^\delta \equiv hg^{\hat{z}} \pmod{n}$ for $\delta = 0, 1, 2, \dots$

By construction, we have $t = k + cz$ and so we obtain $\delta = \lfloor t/c \rfloor - z = \lfloor k/c \rfloor$. If we view function H as a random oracle then, with probability $1/2$, the most significant bit of c is 1. Therefore, with probability at least $1/2$, it follows that

$$\delta = \left\lfloor \frac{k}{c} \right\rfloor \leq \frac{k}{c} < \frac{2^\ell}{2^{\ell-1}} = 2$$

or equivalently that δ is equal to 0 or 1.

Once $z = \hat{z} - \delta$ is known, it is easy to forge the signature on a *chosen* message m' as

$$\sigma' = (t', y', e) \quad \text{with} \quad \begin{cases} y' = yg^{-a} \pmod{N} \\ c' = H(\text{pk}, y', m') \\ t' = t + (c' - c)z + ae \end{cases}, \tag{1}$$

for an arbitrary integer a . It is easy to see that $y'^e g^{t'} h^{c'} \equiv y^e g^{-ae} g^{t+(c-c)z+ae} g^{-zc} \equiv y^e g^{t'} h^{c'} \equiv x \pmod{N}$.

A closer look at the security proof in [16] shows that exponent e is assumed to be $\neq 1$. A signature with $e = 1$ is however valid as it is an odd value (e is not required to be prime in the verification algorithm). This may allow to mount further signature forgeries.

4 Modifying the Scheme

As usual, the security proof given in [16] is by contradiction. It is assumed that there exists a polynomial-time adversary \mathcal{A} making q_S adaptive signature queries on chosen messages and then producing a forgery with non-negligible probability. The forgery is then used to solve some intractable problem — the strong RSA problem.

Definition 1 (Strong RSA problem [1, 10]). *Being given an RSA modulus N and an element $s \in \mathbb{Z}_N$, the strong RSA problem consists in finding $r \in \mathbb{Z}_{>1}$ and $u \in \mathbb{Z}_N$ such that $u^r \equiv s \pmod{N}$.*

For $i \in \{1, \dots, q_S\}$, let m_i be the i^{th} signed message and let $\sigma_i = (t_i, y_i, e_i)$ be the i^{th} signature. Let $\sigma_* = (t_*, y_*, e_*)$ be the forgery on message m_* , produced by \mathcal{A} . Two types of forgeries are distinguished in [16]:

Type I: $e_* = e_j$ for some $j \in \{1, \dots, q_S\}$;

Type II: $e_* \neq e_j$ for all $j \in \{1, \dots, q_S\}$.

For the Type I forger, letting $r := e_* = e_j$ and $c_* = H(\mathbf{pk}, y_*, m_*)$, the authors of [16] obtain the relation

$$\left(\frac{y_*}{y_j}\right)^r \equiv g^{t_j-t} h^{c_j-c} \equiv s^{2(t_j-t - z(c_j-c))} \prod_{i \neq j} e_i \pmod{N} . \tag{2}$$

Next they incorrectly assume that

(A1) $t_j - t_* - z(c_j - c_*)$ (in absolute value) is of length ℓ bits, and

(A2) $t_j - t_* - z(c_j - c_*) \neq 0$,

and hence deduce that the probability of having

$$\gcd\left(r, 2(t_j - t_* - z(c_j - c_*)) \prod_{i \neq j} e_i\right) = \gcd(r, t_j - t_* - z(c_j - c_*)) = 1$$

is overwhelming. As a result, the extended Euclidean algorithm yields integers α and β satisfying $\alpha r + \beta 2(t_j - t_* - z(c_j - c_*)) \prod_{i \neq j} e_i = 1$ and (r, u) with $u := s^\alpha (y_*/y_j)^\beta \pmod{N}$ is a solution to the strong RSA problem.

Remark that the implications of Assumptions (A1) and (A2) are illustrated by the cases $a \neq 0$ and $a = 0$ in the forgeries described in the previous section (cf. Eq. (1)), respectively.

Assumption (A1) should be satisfied by proper parameter definitions and appropriate range verifications in the verification algorithm. Assumption (A2) is more intricate to handle. If $t_j - t_* - z(c_j - c_*) = 0$ then, from Eq. (2), we get $(y_*/y_j)^r \equiv 1 \pmod{N}$ and thus $y_* = y_j$ since $\gcd(r, 2p'q') = 1$. We therefore subdivide the Type I forger into:

Type Ia: $e_* = e_j$ and $y_* \neq y_j$ for some $j \in \{1, \dots, q_S\}$;

Type Ib: $e_* = e_j$ and $y_* = y_j$ for some $j \in \{1, \dots, q_S\}$.

It remains to get a scheme secure in the case corresponding to the Type Ib forger (only Type Ia and Type II are covered in [16]).

The presence of y in the definition of $c = H(\mathbf{pk}, y, m)$ in the original scheme (cf. Fig. 1) seems to indicate that we need to additionally rely on the random oracle model [2]. However, inspecting the security proof given in [16], we see that y is unnecessary in the definition of c for proving the security against Type Ia or Type II forger. We therefore decide to remove it.

Standard techniques for proving the security against the Type Ib forger would lead to defining larger parameters in the signing algorithm. To avoid this, we suggest to exchange the roles of k and t in the original signature scheme. The signature $\sigma = (k, y, e)$ on a message m is then given by

$$y = (xg^{-t})^{1/e} \bmod N \quad \text{with} \quad \begin{cases} t = k - cz \\ c = H(\text{pk}, m) \end{cases} .$$

In more detail, we get the scheme depicted in Fig. 2.

Global parameters. Let ℓ_N, ℓ_H, ℓ_E and ℓ_K be four security parameters, satisfying

$$\ell_E \geq \ell_H + 2 \quad \text{and} \quad \ell_K \gg \ell_N + \ell_H .$$

(Typically $\ell_N = 1024, \ell_H = 160, \ell_E = 162$, and $\ell_K = 1344$.)
 Let also a collision-resistant hash function $H : \{0, 1\}^* \rightarrow \{0, 1\}^{\ell_H}$.

Key generation. Choose two random primes $p = 2p' + 1$ and $q = 2q' + 1$, where p' and q' are primes of equal length, so that $N = pq$ is of length exactly ℓ_N . Choose at random two quadratic residues g and x in \mathbb{Z}_N^* . Finally, for a random integer $z \bmod p'q'$, compute $h = g^{-z} \bmod N$.
 The public key is $\text{pk} = \{g, h, x, N\}$ and the private key is $\text{sk} = \{p, q, z\}$.

Signing. Let $m \in \{0, 1\}^*$ denote the message being signed. Randomly pick an ℓ_K -bit integer t and an ℓ_E -bit prime e . Next compute

$$y = (xg^{-t})^{1/e} \bmod N, \quad c = H(\text{pk}, m), \quad k = t + cz .$$

The signature on message m is $\sigma = (k, y, e)$.

Verification. Signature $\sigma = (k, y, e)$ on message m is accepted iff e is an odd ℓ_E -bit integer and

$$y^e g^k h^c \equiv x \pmod{N}$$

with $c' = H(\text{pk}, m)$.

Fig. 2. Modified TYS signature scheme

It is interesting to note that the scheme we obtained is a disguised version of the Camenisch-Lysyanskaya signature scheme [4]. Actually, it can be seen as an efficient off-line/on-line variant of it. See Appendix A.

5 Conclusion

We have shown that the TYS signature scheme does not meet the EUF-CMA security level. We have mounted an attack against it, yielding part of the secret key. This partial information was then used to produce universal signature forgeries. In a second part, we analyzed why the security was flawed and suggested simple modifications, leading to an efficient off-line/on-line variant of the Camenisch-Lysyanskaya signature scheme.

References

1. N. Barić and B. Pfitzmann, *Collision-free accumulators and fail-stop signature schemes without trees*, Advances in Cryptology – EUROCRYPT '97, Lecture Notes in Computer Science, vol. 1233, Springer-Verlag, 1997, pp. 480–494.
2. M. Bellare and P. Rogaway, *Random oracles are practical: A paradigm for designing efficient protocols*, 1st ACM Conference on Computer and Communications Security, ACM Press, 1993, pp. 62–73.
3. D. Boneh and X. Boyen, *Short signatures without random oracles*, Advances in Cryptology – EUROCRYPT 2004, Lecture Notes in Computer Science, vol. 3027, Springer-Verlag, 2004, pp. 62–73.
4. J. Camenisch and A. Lysyanskaya, *A signature scheme with efficient protocols*, Security in Communication Networks (SCN '02), Lecture Notes in Computer Science, vol. 2576, Springer-Verlag, 2003, pp. 268–289.
5. R. Cramer and I. Damgård, *New generation of secure and practical RSA-based signatures*, Advances in Cryptology – CRYPTO '96, Lecture Notes in Computer Science, vol. 1109, Springer-Verlag, 1996, pp. 173–185.
6. R. Cramer and V. Shoup, *Signature schemes based on the strong RSA assumption*, ACM Transactions on Information and System Security **3** (2000), no. 3, 161–185.
7. C. Dwork and M. Naor, *An efficient existentially unforgeable signature scheme and its applications*, Journal of Cryptology **11** (1998), no. 3, 187–208.
8. A. Fiat and A. Shamir, *How to prove yourself: Practical solutions to identification and signature problems*, Advances in Cryptology – CRYPTO '86 (A.M. Odlyzko, ed.), Lecture Notes in Computer Science, vol. 263, Springer-Verlag, 1987, pp. 186–194.
9. M. Fischlin, *The Cramer-Shoup strong RSA scheme revisited*, Public Key Cryptography – PKC 2003, Lecture Notes in Computer Science, Springer-Verlag, 2003, pp. 116–129.
10. E. Fujisaki and T. Okamoto, *Statistical zero-knowledge protocols to prove modular polynomial equations*, Advances in Cryptology – CRYPTO '97, Lecture Notes in Computer Science, vol. 1294, Springer-Verlag, 1997, pp. 16–30.
11. R. Gennaro, S. Halevi, and T. Rabin, *Secure hash-and-sign signatures without the random oracle*, Advances in Cryptology – EUROCRYPT '99, Lecture Notes in Computer Science, vol. 1592, Springer-Verlag, 1999, pp. 123–139.
12. S. Goldwasser, S. Micali, and R. Rivest, *A digital signature scheme secure against adaptive chosen message attacks*, SIAM Journal of Computing **17** (1988), no. 2, 281–308.
13. G. Poupard and J. Stern, *On the fly signatures based on factoring*, 7th ACM Conference on Computer and Communications Security, ACM Press, 1999, pp. 37–45.
14. A. Shamir and Y. Tauman, *Improved online/offline signature schemes*, Advances in Cryptology – CRYPTO 2001 (J. Kilian, ed.), Lecture Notes in Computer Science, Springer-Verlag, 2001, pp. 355–367.
15. J. Stern, D. Pointcheval, J. Malone-Lee, and N.P. Smart, *Flaws in applying proof methodologies to signature schemes*, Advances in Cryptology – CRYPTO 2002, Lecture Notes in Computer Science, vol. 2442, Springer-Verlag, 2002, pp. 93–110.
16. C.H. Tan, X. Yi, and C.K. Siew, *A new provably secure signature scheme*, IEICE Trans. Fundamentals **E86-A** (2003), no. 10, 2633–2635.
17. H. Zhu, *New digital signature scheme attaining immunity against adaptive chosen message attack*, Chinese Journal of Electronics **10** (2001), no. 4, 484–486.
18. H. Zhu, *A formal proof of Zhu's signature scheme*, Cryptology ePrint Archive, Report 2003/155, 2003.

A Camenisch-Lysyanskaya Signature Scheme

For the reader's convenience, we review below the (single-message) Camenisch-Lysyanskaya signature scheme. We use the same notations as in [4].

Global Parameters. Let $\ell_n, \ell_m, \ell_e, \ell_s$ and ℓ be five security parameters, satisfying

$$\ell_e \geq \ell_m + 2 \quad \text{and} \quad \ell_s = \ell_n + \ell_m + \ell .$$

Key Generation. Choose two random primes $p = 2p' + 1$ and $q = 2q' + 1$, where p' and q' are primes of equal length, so that $n = pq$ is of length exactly ℓ_n . Choose at random three quadratic residues $a, b, c \in \mathbb{Z}_n^*$.

The public key is $\text{pk} = \{n, a, b, c\}$ and the private key is $\text{sk} = \{p, q\}$.

Signing. Let $m \in \{0, 1\}^{\ell_m}$ denote the message being signed. Randomly pick an ℓ_s -bit integer s and an ℓ_e -bit prime e . Next compute

$$v = (a^m b^s c)^{1/e} \pmod n .$$

The signature on message m is $\sigma = (s, v, e)$.

Verification. Signature $\sigma = (s, v, e)$ on message m is accepted iff $v^e \equiv a^m b^s c \pmod n$ and $2^{\ell_e} > e > 2^{\ell_e-1}$.

Remark that the change of variables

$$\left\{ \begin{array}{l} n \leftarrow N \\ c \leftarrow x \\ b \leftarrow g^{-1} \\ a \leftarrow h^{-1} \\ s \leftarrow k \\ m \leftarrow c \\ v \leftarrow y \end{array} \right.$$

transforms the Camenisch-Lysyanskaya signature $(s, v = (a^m b^s c)^{1/e} \pmod n, e)$ into our modified TYS signature (k, y, e) with

$$\begin{aligned} y &= ((h^{-1})^c (g^{-1})^k x)^{1/e} \pmod N \\ &= (g^{zc-k} x)^{1/e} \pmod N \\ &= (xg^{-t})^{1/e} \pmod N \end{aligned}$$

where $t = k - cz$.

Efficient Partially Blind Signatures with Provable Security

Qianhong Wu¹, Willy Susilo¹, Yi Mu¹, and Fanguo Zhang²

¹ School of Information Technology and Computer Science,
University of Wollongong, Wollongong NSW 2522, Australia
{qhw, wsusilo, ymu}@uow.edu.au

² School of Information Science and Technology,
Sun Yat-sen University, Guangzhou 510275,
Guangdong Province, P.R. China
isdzhfg@zsu.edu.cn

Abstract. Blind signatures play a central role in applications such as e-cash and e-voting systems. The notion of partially blind signature is a more applicable variant such that the part of the message contains some common information pre-agreed by the signer and the signature requester in an unblinded form. In this paper, we propose two efficient partially blind signatures with provable security in the random oracle model. The former is based on witness indistinguishable (WI) signatures. Compared with the state-of-the-art construction due to Abe and Fujisaki [1], our scheme is 25% more efficient while enjoys the same level of security. The latter is a partially blind Schnorr signature without relying on witness indistinguishability. It enjoys the same level of security and efficiency as the underlying blind signature.

1 Introduction

The notion of blind signature was first introduced by Chaum in [4]. After Chaum's first scheme based on RSA, some discrete-log based signature schemes were converted into blind signatures (e.g., [6], [8], [10]). Chaum remarked that the notion is the only way to implement electronic cash simulating the ones in reality. However, when the original notion of blind signatures is implemented for e-cash systems, as the bank knows nothing about the resulting signature or the signed message, some problems are posed, such as how to prevent misuse of the signature or how to embed the information such as the issuing date, expiration date, face value of e-cash and so on. There are two conventional solutions. The first one is that the bank uses different public keys to link with such common information. In this case, the shops and customers must always carry a list of those public keys in their electronic wallet, which is typically a smart card whose memory is very limited. The second solution is that the bank can use the cut-and-choose algorithm [4] in the withdrawal phase. This solution is also very inefficient.

To address such issues in applications, partially blind signatures were introduced by Abe and Fujisaki [1] to allow the signer to explicitly include some

pre-agreed information in blind signatures. Using partially blind signatures in e-cash systems, the bank can be relieved from maintaining a unlimitedly growing database. The bank assures that each e-cash contains the information it desires, such as the date information. By embedding an expiration date into each e-cash issued by the bank, all expired e-cash recorded in the banks database can be removed. At the same time, since face values are embedded into e-cash, the bank knows the value on each e-cash blindly issued. The notion attracts a lot of attentions and has been extensively implemented under different assumptions (e.g., [1], [2], [7], [18], [5], [21]).

In this paper, we propose two efficient partially blind signatures with provable security in the random oracle model. The former is based on witness indistinguishable signatures [3]. Compared with the state-of-the-art construction due to Abe and Fujisaki [1], our scheme is 25% more efficient while enjoys the same level of security. The latter is a partially blind Schnorr signature without relying on witness indistinguishability. The merit is that, after an efficient publicly available evolution of public keys, we achieve partially blindness from blind signatures *without* introducing additional overhead or degrading the security of the underlying schemes.

2 Notations and Definitions

2.1 Notations

A negligible function is a function $\varepsilon(\lambda)$ such that for all polynomials $poly(\lambda)$, $1/\varepsilon(\lambda) < 1/poly(\lambda)$ holds for all sufficient large λ . PPT stands for *probabilistic polynomial-time*. $KS\{x : y = f(x)\}(m)$ represents a knowledge signature on message m that is transformed from the zero-knowledge proof of a secret value x satisfying $y = f(x)$ with the well-known Fiat-Shamir transformation. An efficient algorithm $\mathcal{A}(\cdot)$ is a probabilistic Turing machine running in expected polynomial time. An adversary \mathcal{A} is a PPT interactive Turing machine. If $\mathcal{A}(\cdot)$ is an efficient algorithm and x is an input for \mathcal{A} , then $\mathcal{A}(x)$ denotes the probability space that assigns to a string σ the probability that \mathcal{A} , on input x , outputs σ . An efficient algorithm is deterministic if for every input x , the probability mass of $\mathcal{A}(x)$ is concentrated on a signed output string σ . For a probability space \mathbb{P} , $x \leftarrow \mathbb{P}$ denotes the algorithm that samples a random element according to \mathbb{P} . For a finite set \mathbb{X} , $x \leftarrow \mathbb{X}$ denotes the algorithm that samples an element uniformly at random from \mathbb{X} .

2.2 Partially Blind Signatures

We review the security definitions of partially blind signatures in [2] which follows from [9]. They are similar to those of conventional blind signature.

A partially blind signature is a tuple $PBS = (\mathcal{G}^{PBS}(\cdot), Sig^{PBS}(\cdot), Ver^{PBS}(\cdot))$:

- $(sk, pk) \leftarrow \mathcal{G}^{PBS}(\mathbf{1}^\lambda)$ is a PPT algorithm which, on input a security parameter λ , outputs the signer's private/public key pair (sk, pk) .

- $\sigma \leftarrow \text{Sig}_{sk;m}^{PBS}(pk, \text{info})$ is a two-party protocol between a signer \mathcal{S} and a user \mathcal{U} . They have common input pk , which is the public key, and a pre-agreed common information info between the signer and the user. \mathcal{S} has private input sk , which is its secret key, and its inner random coin flips. \mathcal{U} has private input m , which is the message to be signed, and its inner random coin flips. At the end of the interaction between the two parties, \mathcal{S} outputs one of the two messages: *completed*, *not – completed*, and \mathcal{U} outputs either *fail* or a signature σ on m .
- $0/1 \leftarrow \text{Ver}^{PBS}(m, \text{info}, \sigma, pk)$ is a polynomial-time deterministic algorithm which, on input a message-signature pair (m, σ) , pre-agreed information info and the signer's public key pk , returns 1 or 0 for *accept* or *reject*, respectively. If *accept*, the message-signature pair is valid.

2.3 Security of Partially Blind Signatures

Similarly to blind signatures, the security of partially blind signature schemes has three critical aspects: correctness, partial blindness and $(l, l + 1)$ -unforgeability. They are formally defined as follows.

Definition 1 (Correctness). *If the signer \mathcal{S} and the user \mathcal{U} honestly follow the protocol $\text{Sig}^{PBS}(\cdot)$, \mathcal{U} will output a signature σ on m accepted by $\text{Ver}^{PBS}(\cdot)$ with a dominant probability.*

Definition 2 (Partial Blindness). *The partial blindness is defined via an experiment involving an adversarial signer \mathcal{A} . The experiment is parameterized by a bit b and security parameter λ . First $\mathcal{G}^{PBS}(1^\lambda)$ is correctly run to generate the adversarial signer \mathcal{A} 's secret key/public key pair (sk, pk) . Then, \mathcal{A} outputs a pair of messages (m_0, m_1) lexicographically ordered. In the next stage of the experiment \mathcal{A} engages in two (possibly correlated and interleaved) runs with two honest users, with inputs m_b and m_{1-b} , respectively. If both users obtain valid signatures, on their respective message, \mathcal{A} is also given these two signatures; otherwise there is no extra input to \mathcal{A} ; in either case, \mathcal{A} is required to output a bit b' . The advantage of \mathcal{A} is defined by:*

$$\text{Adv}_{PBS, \mathcal{A}}^{\text{Bld}}(\lambda) = 2\Pr[b = b'] - 1.$$

A PBS scheme satisfies partial blindness, if for any PPT adversary signer \mathcal{A} , the function $\text{Adv}_{PBS, \text{Ad}}^{\text{Bld}}(\lambda)$ is negligible in λ . If \mathcal{A} 's computational power is unlimited, the blindness is unconditional.

Definition 3 ($(l, l+1)$ -Unforgeability). *The $(l, l+1)$ -Unforgeability of a partially blind signature scheme is defined via an experiment parameterized by security parameters l and λ . The experiment involves an adversarial user \mathcal{A} and is as follows: First $\mathcal{G}^{PBS}(1^\lambda)$ is run to generate the signer's secret/public key pair (sk, pk) . The signer and the adversarial user \mathcal{A} also negotiate with the common information info . Then, the adversary user \mathcal{A} engages in polynomially many (possibly correlated and interleaved) runs of the protocol with the signer. Finally*

\mathcal{A} outputs a list of message-signature pairs $\{(m_1, \sigma_1), (m_2, \sigma_2), \dots, (m_t, \sigma_t)\}$ with $m_i \neq m_j$ and a string \mathbf{info}' . Let l be the number of runs successfully completed by the signer. Define the advantage of \mathcal{A}

$$Adv_{PBS, \mathcal{A}}^{Unf}(\lambda) = \Pr[\forall 1 \leq i \leq t, Ver^{PBS}(m_i, \mathbf{info}, \sigma_i, pk) = 1 \wedge (l < t)] + \Pr[\exists 1 \leq i \leq t, Ver^{PBS}(m_i, \mathbf{info}', \sigma_i, pk) = 1 \wedge \mathbf{info}' \neq \mathbf{info}]$$

and say that the PBS scheme is $(l, l+1)$ -unforgeable if $Adv_{PBS, \mathcal{A}}^{Unf}(\lambda)$ is negligible for any PPT adversary user \mathcal{A} .

3 Proposed Partially Blind Signatures

In this section, we propose two partially blind signatures. The first one is a blind version of the knowledge proof signature that the signer knows a discrete logarithm of Schnorr public key y or the hashed value of common information c to a given base g . So we name it *partially blind OR-Schnorr signature*. It is similar to the construction in [2] but more efficient. The second one is a blind version of knowledge proof signature where the signer knows $(x + F(c))^{-1}$ which is the inverse of the sum of Schnorr secret key and the hashed value of common information c . Hence, we refer to it as partially blind inverse-Schnorr signature. It can be generalized to suit other signatures derived from knowledge proofs of discrete logarithms.

3.1 Partially Blind OR-Schnorr Signature

Let G be a cyclic group with prime order q , and g an element in G whose order is q . We assume that any polynomial-time algorithm solves $\log_g h$ in \mathbb{Z}_q only with negligible probability when h is selected randomly from G . Let $H : \{0, 1\}^* \rightarrow \mathbb{Z}_q$ and $F : \{0, 1\}^* \rightarrow G$ be public cryptographic hash functions. Let $x \in \mathbb{Z}_q$ be a secret key and $y = g^x$ be the corresponding public key. The signer and the user first agree on common information c in a predetermined way. Then, they execute the signature issuing protocol on the user's blind message m as follows.

- (Initialization). The signer randomly selects $r, u \leftarrow \mathbb{Z}_q^*$, and computes $z = F(c)$, $a = g^r z^u$. The signer sends a to the user as a commitment.
- (Blinding). The user randomly selects $t_1, t_2, t_3 \leftarrow \mathbb{Z}_q$ as blind factors, and computes $z = F(c)$, $\alpha = ag^{t_1} y^{t_2} z^{t_3}$, $\varepsilon = H(\alpha || c || z || m)$, $e = \varepsilon - t_2 - t_3 \pmod q$. The user sends e to the signer.
- (Signing). The signer sends back v, s to the user, where $v = e - u \pmod q$, $s = r - vx \pmod q$.
- (Unblinding). The user computes $\sigma = s + t_1 \pmod q$, $\rho = v + t_2 \pmod q$, $\delta = e - v + t_3 \pmod q$. It outputs (σ, ρ, δ) as the resulting signature on the message m and the pre-agreed common information c .
- (Verification). The signature is valid if and only if $\rho + \delta = H(g^\sigma y^\rho z^\delta || c || z || m)$ and $z = F(c)$.

Now we consider the security of the scheme. Note that, in the Signing step, the signer returns $v = e - u$ and the user knows e . So the user can compute u ; that is, the user learns the component u of the indistinguishable witness (r, u) such that $a = g^r z^u$. However, due to the witness indistinguishability, u can be replaced by any other value u' in the random oracle model. This is indeed the key point that the signer procedure can be simulated without knowledge of the secret key x satisfying $y = g^x$ in the random oracle model. The detailed simulation is provided in Theorem 3.

Theorem 1. (*Correctness*) *If the signer and user follows the protocol, the output of the user will be accepted by the verification algorithm.*

Proof. Note that $g^\sigma y^\rho z^\delta = g^{s+t_1} y^{v+t_2} z^{e-v+t_3} = g^s y^v z^{e-v} g^{t_1} y^{t_2} z^{t_3} = g^{r-vx} y^v z^u g^{t_1} y^{t_2} z^{t_3} = a g^{t_1} y^{t_2} z^{t_3} = \alpha$ and $\rho + \delta = u + v + t_2 + t_3 = e + t_2 + t_3 = \varepsilon$. It follows that $\rho + \delta = H(g^\sigma y^\rho z^\delta || c || z || m)$, $z = F(c)$. The verification holds.

Theorem 2. (*Partial Blindness*) *The above partially blind signature is unconditionally blind.*

Proof. It is sufficient to prove that, for any view (a, e, v, s, c) of the adversary signer \mathcal{S}^* and any signature pair $(\sigma, \rho, \delta, m, c)$, there exists a blind factor tuple that maps the view and the signature pair.

For $i = 0, 1$, let (a_i, e_i, v_i, s_i, c) be views of \mathcal{S}^* and $(\sigma_j, \rho_j, \delta_j, m_j, c)$ two valid partially blind signatures from user $j = 0, 1$, respectively. Let $t_1 = \sigma_j - s_i$, $t_2 = \rho_j - v_i$, $t_3 = \delta_j - e_i + v_i$. Since $a_i = g^{r_i} z^{u_i}$ and the signatures are valid, it follows that

$$\begin{aligned} \rho_j + \delta_j &= H(g^{\sigma_j} y^{\rho_j} z^{\delta_j} || z || m_j) \\ &= H(a_i g^{-r_i} z^{-u_i} g^{\sigma_j} y^{\rho_j} z^{\delta_j} || z || m_j) = H(a_i g^{\sigma_j - r_i} z^{\delta_j - u_i} y^{\rho_j} || z || m_j) \\ &= H(a_i g^{\sigma_j - s_i - v_i} z^{\delta_j - e_i + v_i} y^{\rho_j} || z || m_j) = H(a_i g^{\sigma_j - s_i} z^{\delta_j - e_i + v_i} y^{\rho_j - v_i} || z || m_j) \\ &= H(a_i g^{t_1} y^{t_2} z^{t_3} || z || m_j). \end{aligned}$$

Hence, given $j \in \{0, 1\}$, (a_0, e_0, v_0, s_0, c) and (a_1, e_1, v_1, s_1, c) have the same relation with $(\sigma_j, \rho_j, \delta_j, m_j, c)$ defined by the signing protocol. Therefore, given a signature $(\sigma_j, \rho_j, \delta_j, m_j, c)$, an infinitely powerful \mathcal{S}^* can guess j correctly with probability exactly $1/2$.

Theorem 3. (*Unforgeability*) *In the random oracle model, under the DLP assumption, the above partially blind signature is $(l, l + 1)$ -unforgeable against the sequential attack if $l < \text{poly}(\log \log q)$.*

Proof. The proof is similar to that in [2] except the knowledge extraction part. We first prove that after l interactions with the signer, the adversary user \mathcal{U}^* can not forge a valid signature that the common information has never appeared in the interactions. This proof follows that used in [11]. By using \mathcal{U}^* , an interactive algorithm Sim is constructed to forge the plain version (without blindness) of the above scheme. Then, we use Sim to break the DLP (Discrete Logarithm Problem) assumption.

Let Q_F, Q_H denote the maximum number of queries asked from U^* to F and H , Q_S the maximum number of invocation of signer \mathcal{S} . F and H reply with the same values for the duplicated queries. Given the DLP-challenge (G, q, g, y) , we are required to find $x = \log_g y \in \mathbb{Z}_q$. Sim simulates F, H, S as follows. Randomly select $I \leftarrow \{1, \dots, Q_F + Q_S\}$ and $J \leftarrow \{1, \dots, Q_H + Q_S\}$. For the i -th query, if $i = I$, return $z = F(c_I)$; else return $z = g^{\varpi_i}$, where $\varpi_i \in \mathbb{Z}_q$. For the j -th query to H , forward to H if $j = J$; else return a random integer in \mathbb{Z}_q . For query to \mathcal{S} , if common information $c \neq c_I$, simulate the signature by using witness ϖ_i , else the simulation fails. If U^* outputs valid signature pair $(\sigma, \rho, \delta, m, c)$, Sim outputs it. Since F and H may be asked during the simulation of the signer \mathcal{S} , it may have at most $Q_F + Q_S$ and $Q_H + Q_S$ points. The simulation of \mathcal{S} is perfect if $c \neq c_I$ due to witness indistinguishability. Without any query to F, H , the adversary can successfully output a valid signature with negligible probability in $\log q$. Suppose that the success probability of U^* is at least ϵ_U , then the success probability of Sim is at least $\epsilon_{\text{sim}} = (Q_F + Q_S - 1)\epsilon_U / (Q_F + Q_S)^2(Q_H + Q_S)^2$.

Now we construct an algorithm to solve $x = \log_g y \in \mathbb{Z}_q$ by using Sim as a subroutine. In this case, if Sim query to F , a random integer $\vartheta \in \mathbb{Z}_q$ is selected and yg^ϑ is returned to Sim . Note that query to F is before query to H . Now we use the standard rewind technique. After $1/\epsilon_{\text{sim}}$ trials, Sim will output a valid signature (σ, ρ, δ) with probability at least $1 - e^{-1}$ (here e is the base of natural logarithms). By repeating this rewind-trial $2/\epsilon_{\text{sim}}$ times, we obtain another valid signature $(\sigma', \rho', \delta')$, with probability at least $(1 - e^{-1})/2$. Hence, with a constant probability and polynomial running time, we have two valid signatures with the same first transcript a . It follows that $\sigma + x\rho + (x + \vartheta)\delta = \sigma' + x\rho' + (x + \vartheta)\delta'$ and $\rho + \delta \neq \rho' + \delta'$. Hence, we extract that $x = (\sigma - \sigma' + \vartheta(\delta - \delta')) / (\rho + \delta - \rho' - \delta')$.

Then we consider an adversary U^* which outputs $l + 1$ valid signature with the same common information c . Similarly, we will construct an algorithm Sim to solve the DLP assumption by using U^* as a black box. Assume that a DLP-challenge (G, q, g, z) is given to Sim . First, Sim selects $b \in_U \{0, 1\}$ and sets $(y, z) = (g^x, zg^\vartheta)$ if $b = 0$, or $(y, z) = (zg^\vartheta, g^w)$ if $b = 1$ by choosing ϑ and x (or w) randomly from \mathbb{Z}_q . F is defined so that it returns an appropriate value of z according to the choice. We assume that $b = 0$ is chosen and $(y, z) = (g^x, zg^\vartheta)$ is set. Sim can then simulate signer \mathcal{S} , since the protocol between \mathcal{S} and U^* is witness indistinguishable and $x = \log_g y$ is sufficient for Sim to complete the protocol. Let \mathcal{S}' denote the signer simulated by Sim . If U^* is successful with probability at least ϵ_U , we can find a random tape string for U^* and \mathcal{S}' with probability at least $1/2$ such that \mathcal{S}' succeeds with probability at least $\epsilon_U / 2$. By employing U^* as a black-box, we can construct U' , which has exactly the same interface with \mathcal{S}' as U^* has, and plays the role of an impersonator in the interactive identification protocol with the verifier. When U^* asks at most q_F queries to random oracle H , U' is successful in completing the identification protocol with the verification procedure with probability at least $\epsilon_U / 2Q_H^{l+1}$, since, with probability greater than $1/2Q_H^{l+1}$, U' can guess a correct selection of $l + 1$ queries that U' eventually uses in the forgery. Sim then use the standard rewind technique for an interactive protocol to compute the discrete logarithm.

Sim first runs \mathcal{U}' with \mathcal{S}' and the verifier, and find a successful challenge tuple $(\varepsilon_1, \dots, \varepsilon_{l+1})$. Sim then randomly chooses an index, $i \in \{1, \dots, l + 1\}$, and replays with the same environments and random tapes except different challenge tuple $(\varepsilon_1, \dots, \varepsilon_{i-1}, \varepsilon'_i, \dots, \varepsilon'_{l+1})$ where the first $i - 1$ challenges are unchanged. It follows that $\sigma_i + x\rho_i + (w_0 + \vartheta)\delta = \sigma'_i + x\rho'_i + (w_0 + \vartheta)\delta'_i$ and $\rho_i + \delta_i \neq \rho'_i + \delta'_i$. If $\delta_i \neq \delta'_i$, we can extract that $w_0 = (\sigma_i - \sigma'_i + x(\rho_i - \rho'_i))/(\delta'_i - \delta_i) - \vartheta \bmod q$. Similarly, in the case of $b = 1$, if $\rho_i \neq \rho'_i$, we can extract $w_0 = (\sigma_i - \sigma'_i + w(\delta_i - \delta'_i))/(\rho'_i - \rho_i) - \vartheta \bmod q$. Let the event $\delta_i \neq \delta'_i$ happen with probability P . Since $\rho_i + \delta_i \neq \rho'_i + \delta'_i$, the event $\rho_i \neq \rho'_i$ happens with probability at least $1 - P$. Therefore, after finding a collision in the rewinding, we can extract w_0 with probability at least $0.5P + 0.5(1 - P) = 1/2$ and solve the DLP-challenge. The remaining part to evaluate the collusion probability is the same as that in [2] and hence, it is omitted.

3.2 Partially Blind Inverse-Schnorr Signatures

As the above scheme, most DLP-based partially blind signatures exploit the witness indistinguishable signatures. In the following, we exploit the simplest solution to transform blind signatures to partially blind signatures by linking different public keys with different common information. However, we must address the issue to generate and manage the exponentially many public/private key pairs corresponding to the different common information. The trick is to use a publicly available deterministic algorithm to evolve the public key and enable the signer to evolve its private key accordingly. To illustrate this technique, we implement this transformation with the known blind Schnorr signature [22].

Let us assume that the same settings as the above scheme are used. Let G be a cyclic group with prime order q , and g a generator of G . $H, F : \{0, 1\}^* \rightarrow \mathbb{Z}_q$ are public cryptographic hash functions. $y = g^x$ is the signer's public key and $x \in \mathbb{Z}_q$ is the corresponding secret key. Assume that the signer and the user agree on common information c . The resulting signature is a blind version of the knowledge signature $KS\{(x + F(c))^{-1} | g = (yg^{F(c)})^{(x + F(c))^{-1}}\}(m)$. The signature issuing protocol on the user's blind message m is as follows.

- (Key Evolution.) The signer computes $z = F(c)$, $Y = yg^z$ as its new public key and set $X = (x + z)^{-1} \bmod q$ accordingly as its new private key.
- (Initialization.) The signer randomly selects $r \leftarrow \mathbb{Z}_q$, and sends $a = Y^r$ to the user as a commitment.
- (Blinding.) The user computes $z = F(c)$, $Y = yg^z$. It picks random numbers $\alpha, \beta \leftarrow \mathbb{Z}_q$, computes $b = aY^\alpha g^\beta$, $\varepsilon = H(b || z || m)$, and returns a challenge $d = \varepsilon - \beta \bmod q$.
- (Signing.) The signer sends back $w = r - dX \bmod q$.
- (Unblinding.) The user computes $s = w + \alpha \bmod q$ and outputs (ε, s) as the resulting signature on m , and the pre-agreed common information c .
- (Verification.) The signature is valid if and only if $\varepsilon = H(Y^s g^\varepsilon || z || m)$ and $z = F(c), Y = yg^z$.

Clearly, except for an efficient additional key evolution procedure with c as the common information, the scheme is the same as the blind Schnorr signature.

Hence, it enjoys the same security as the blind Schnorr signature. Since the blind Schnorr signature is not witness indistinguishable, its security is not proven with the standard DLP-assumption. Under the ROS assumption [15] (or see Appendix A), Schnorr proved that the scheme is secure in the random oracle model plus generic group model.

Theorem 4. *The above partially blind inverse-Schnorr Signature is correct and unconditionally partially blind, if the ROS assumption holds in the random oracle model plus the generic group model, the above partially blind signature is $(l, l+1)$ -unforgeable.*

Proof. The correctness and the unconditional partial blindness are directly from those of the underlying blind Schnorr signature. After l interactions with the signer, the adversary \mathcal{U}^* can not forge a valid signature that the common information has never appeared in the interactions. Or by using \mathcal{U}^* , another adversary \mathcal{A} can be constructed to break the plain Schnorr signature (without blindness). The plain Schnorr signature has been proven secure in the random oracle model under the DLP assumption [13]. For an adversary \mathcal{U}^* which outputs $l + 1$ valid signature with the same common information c , the adversary \mathcal{U}^* can be directly used to break $(l, l + 1)$ -unforgeability of the underlying blind Schnorr signature which has been proven secure if the ROS assumption holds in the random oracle model plus the generic group model [15] (The DLP assumption is weaker than the general group model assumption as it has been proven that DLP is difficult in the general group model [19],[20]). Hence, we obtain the above claim.

This transformation can be similarly extended to other DLP-related blind signatures such as the blind versions of Okamoto-Schnorr signature, ElGamal signature, DSA, etc. For instance, for the Okamoto-Schnorr signature with public key $y = g^{x_1}h^{x_2}$ and private keys $x_1, x_2 \in \mathbb{Z}_q$, the corresponding partially blind signature can be similarly built as the blind knowledge signature $KS\{\alpha, \beta|g = Y^\alpha h^\beta\}(m)$, where $Y = yg^z$, $\alpha = (x_1 + z)^{-1}$, $\beta = x_2(x_1 + z)^{-1}$, $z = F(c)$ and c is the pre-agreed common information. Clearly, after the key evolution, the partially blind signatures enjoy the same efficiency of the underlying blind signatures.

3.3 Comparison of Partially Blind Signatures

In this section, we give a comparison between the proposed two schemes (denoted by PBS-OR and PBS-Inv respectively) and the state-of-the-art partially blind signature [2] (denoted by PBS-AO) in terms of efficiency and security.

In the above table, Exp denotes the exponentiation in group G . $\lambda = \log q$. ROM stands for the random oracle model and GM means the generic group model. DLP represents the discrete logarithm problem in group G and ROS denotes the ROS problem introduced by Schnorr [15]. The computation complexity is in terms of exponentiation without any optimization, that is, a two-base exponentiation is calculated as two single-base exponentiations and so on. Clearly, compared with the scheme in [2], our scheme PBS-OR is about 25% more efficient while enjoying the same level of security. The proposed PBS-Inv is the most

Table 1. Comparison of the state-of-the-art partially blind signatures

	Signer	User	Verification	Length: bits	Security
PBS-AO	2 Exp	4 Exp	4 Exp	4λ	ROM+DLP
PBS-OR	2 Exp	3 Exp	3 Exp	3λ	ROM+DLP
PBS-Inv	1 Exp	2 Exp	2 Exp	2λ	ROM+GM+ROS

efficient scheme among the three with a slightly stronger security assumption, i.e., the generic group assumption.

Note that the unforgeability of PBS-OR and PBS-AO is proven against the *sequential attack* formalized in [14]. The PBS-Inv is $(l, l + 1)$ -unforgeable against the *general parallel attack* formalized in [15]. However, it requires an additional ROS assumption. For (partially) blind signatures (including the PBS-AO scheme [2]) derived from knowledge proof of discrete logarithms, to our knowledge, *no* scheme has been proven secure against *generic parallel attack* (which is more powerful than the sequential attack) without relying on the ROS assumption which is a very strong assumption [16]. Although some scheme claimed such security [17], further analysis shows the claimed security is overestimated (Appendix A). On the other hand, these DLP-derived (partially) blind signatures can be proven secure against generic parallel attack under the ROS assumption using the similar techniques due to Schnorr [15]. From the state-of-the-art algorithm to solve the ROS problem [15], it requires subexponential running time $O(2^{2\sqrt{\log q}})$ [16] to break ROS problem which is less than previous evaluation $O(\sqrt{q})$ [15], where q is the order of group G . We must set a more conservative security parameter $q > 2^{1600}$ rather than $q > 2^{160}$.

4 Conclusions

In this paper, efficient partially blind signatures are proposed with provable security. Compared with the scheme due to Abe *et al.*, our partially blind signatures based on witness indistinguishability enjoy the same level of security but are more efficient. We also suggest a simple yet efficient technique to transform blind signatures to partially blind signatures with the same security property as the underlying blind signatures. We implement this transformation with the known blind Schnorr signature.

Because of page limitation, we omit the additional security analysis. We refer the reader to the full version of this paper.

References

1. M. Abe, E. Fujisaki. How to date blind signatures. In *Asiacrypt'96*, LNCS 1163, pp. 244-251. Springer-Verlag, 1996.
2. M. Abe and T. Okamoto. Provably secure partially blind signatures. In *Crypto'00*, LNCS 1880, pp. 271-286. Springer-Verlag, 2000.

3. R. Cramer, I. Damgård, and B. Schoenmakers. Proofs of partial knowledge and simplified design of witness hiding protocols. In *Crypto'94*, LNCS 839, pp. 174-187. Springer-Verlag, 1994.
4. D. Chaum. Blind signatures for untraceable payments. In *Crypto'82*, pp. 199-204. Prenum Publishing Corporation, 1982.
5. S. S. M. Chow, L. C.K. Hui, S.M. Yiu, and K. P. Chow. Two improved partially blind signature schemes from bilinear pairings. In *ACISP'05*, LNCS 3574, pp. 316-328. Springer-Verlag, 2005.
6. J. Camenisch, J.-M. Piveteau, and M. Stadler. Blind signatures based on the discrete logarithm problem. In *Eurocrypt'94*, LNCS 950, pp. 428-432. Springer-Verlag, 1995.
7. C.I. Fan and C.L. Lei, Low-computation partially blind signatures for electronic cash. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* E81- A(5) (1998) 818-824.
8. H. Horster, M. Michels, and H. Petersen. Meta-message recovery and meta-blind signature schemes based on the discrete logarithm problem and their applications. In *Asiacrypt '92*, LNCS 917, pp. 224-237. Springer-Verlag, 1992.
9. A. Juels, M. Luby, and R. Ostrovsky. Security of blind digital signatures. In *Crypto'97*, LNCS 1294, pp. 150-164. Springer-Verlag, 1997.
10. T. Okamoto and K. Ohta. Divertible zero knowledge interactive proofs and commutative random self-reducibility. In *Eurocrypt'89*, LNCS 434, pp. 134-149. Springer-Verlag, 1990.
11. K. Ohta, T. Okamoto. On concrete security treatment of signatures derived from identification. In *Crypto'98*, LNCS 1462, pp. 354-369. Springer-Verlag, 1998.
12. D. Pointcheval. Strengthened security for blind signatures. In *Eurocrypt'98*, LNCS 1403, pp. 391-405. Springer-Verlag, 1998.
13. D. Pointcheval, J. Stern. Security arguments for digital signatures and blind signatures. *Journal of Cryptology*,13(3):361-396, 2000.
14. D. Pointcheval, J. Stern. Provably secure blind signature schemes. In *Asiacrypt'96*, LNCS 1163, pp. 252-265. Springer-Verlag, 1996.
15. C. Schnorr. Security of Blind Discrete Log Signatures against Interactive Attacks. In *ICICS'01*, LNCS 2229, pp. 1-12. Springer-verlag, 2001.
16. D. Wagner. A generalized birthday problem. In *Crypto'02*, LNCS 2442, pp. 288-304, 2002. Springer-Verlag, 2002.
17. F. Zhang, K. Kim. Efficient ID-Based blind signature and proxy signature from bilinear pairings. In *ACISP'03*, LNCS 2727, pp. 312-323. Springer-verlag, 2003.
18. F. Zhang, R. Safavi-Naini, and W. Susilo. Efficient verifiably encrypted signature and partially blind signature from bilinear pairings. In *Indocrypt'03*, LNCS 2904, pp. 191-204. Springer-Verlag, 2003.
19. V. I. Nechaev. Complexity of a determinate algorithm for the discrete logarithm. *Mathematical Notes* 55, pp. 165-172, 1994.
20. V. Shoup. Low bounds for discrete logarithms and related problems. in *Eurocrypt'07*, LNCS 1233, pp. 256-266, Springer-Verlag, 1997.
21. G. Maitland, and C. Boyd. A Provably Secure Restrictive Partially Blind Signature Scheme. in *Public Key Cryptography (PKC 2002)*, LNCS 2274, pp. 991-1014, Springer-Verlag, 2002.
22. T. Okamoto. Provable secure and practice identification schemes and corresponding signature schemes. in *Crypt'92*, LNCS 740, pp. 31-53, Springer-Verlag, 1993.

A Framework for Robust Group Key Agreement

Jens-Matthias Bohli

Institut für Algorithmen und Kognitive Systeme,
Universität Karlsruhe, Germany
bohli@ira.uka.de

Abstract. Considering a protocol of Tseng, we show that a group key agreement protocol that resists attacks by malicious insiders in the authenticated broadcast model, loses this security when it is transferred into an unauthenticated point-to-point network with the protocol compiler introduced by Katz and Yung. We develop a protocol framework that allows to transform passively secure protocols into protocols that provide security against malicious insiders and active adversaries in an unauthenticated point-to-point network and, in contrast to existing protocol compilers, does not increase the number of rounds. Our protocol particularly uses the session identifier to achieve the security. By applying the framework to the Burmester-Desmedt protocol we obtain a new 2 round protocol that is provably secure against active adversaries and malicious participants.

1 Introduction

A group key establishment protocol allows $n \geq 2$ users to agree upon a common secret session key for private communication. One subclass of those protocols are *key agreement* protocols, where all participants contribute to the key in a way such that no collusion of malicious participants can enforce a previously known session key. The first group key establishment protocol was proposed in [12], some protocols that will be considered in this work were introduced in [6]. Unfortunately, no security proof was given for those protocols. A model for provably secure key establishment protocols is introduced in [2] and later adapted for group key establishments in [1, 5, 14], different models are developed in [17, 7, 18]. Katz and Yung [14] also present a protocol compiler that transforms a (non-authenticated) protocol that is secure against passive adversaries into an authenticated protocol that is also secure against active adversaries. A recent overview of the indistinguishability-based models is given in [8].

However, these models did not allow for malicious participants attacking the protocol even though a majority of attacks stem from malicious insiders (cf. [11]). Recently, models that consider malicious participants were developed in [3, 13]. A protocol compiler turning a protocol that is secure against active adversaries in a protocol secure against malicious insiders is given in [13]. A protocol that is secure against malicious insiders in those models is presented in [3]. The protocol takes only two rounds and therefore cannot be obtained by the known protocol compilers.

Protocols that resist insider attacks were previously known in the *authenticated broadcast model*, where the adversary cannot delay or alter messages that are sent via the broadcast channel [15, 16, 21, 20, 19]. However, this model is not compatible with the established proof models mentioned above. In this contribution we show, considering [19] as an example, that the security against malicious insiders does unfortunately not survive a straightforward transformation with the Katz-Yung compiler into a protocol that is secure against active adversaries in the unauthenticated link model.

In this contribution, we develop a framework to transform any 2 round key agreement protocol that offers security against passive adversaries in a protocol that is secure against active adversaries and malicious participants in an unauthenticated network. The framework generalizes the protocols of [19, 3].

2 Preliminaries

We base on the model from [3], that extends the previous proof models [5, 14] which did not consider malicious participants. We give now a short overview of the building blocks of these models.

Initialization, Participants and the Adversary. We assume a fixed set of users $\mathcal{U} = \{U_1, \dots, U_n\}$. In an initialization phase, a key generation algorithm $Gen(1^k)$ is executed to generate a public/private key pair for every user U_i . Every user stores his secret key and the public keys of all other users.

The participants of an execution of the key establishment protocol can be any subset of \mathcal{U} . To model that a user can participate in multiple executions, every user U_i is represented by several instances Π_i^s . All instances Π_i^s of a user have access to the user's secret key and all instances hold an individual set of variables state_i^s , sid_i^s , pid_i^s , sk_i^s , term_i^s , used_i^s and acc_i^s . After a successful protocol run, sid_i^s will uniquely name the session and sk_i^s will store the session key. The users who participated in this session from U_i 's point of view are stored in pid_i^s .

The adversary interacts with the instances via oracles

Execute($\{U_1, U_2, \dots, U_r\}$) Delivers a protocol transcript.

Send(U_i, s_i, M) Sends M to $\Pi_i^{s_i}$ and outputs the instance's reply. A special message is used to assign a role and pid and initiate the protocol.

Reveal(U_i, s_i) Returns the session key $\text{sk}_i^{s_i}$.

Corrupt(U_i) Returns the long term secret key of U_i .

Test(U_i, s_i) Only one query of this form is allowed. If $\text{sk}_i^{s_i}$ is defined, with probability 1/2 the session key $\text{sk}_i^{s_i}$ and with probability 1/2 a uniformly chosen random session key is returned.

Partnering, Freshness and Security Against Active Adversaries. To define security of a key establishment protocol we need to exclude those cases when the adversary trivially knows the established session key. We start by calling two instances $\Pi_i^{s_i}$ and $\Pi_j^{s_j}$ *partnered* if $\text{sid}_i^{s_i} = \text{sid}_j^{s_j}$, $\text{acc}_i^{s_i} = \text{acc}_j^{s_j} = \text{true}$ and both $U_j \in \text{pid}_i^{s_i}$ and $U_i \in \text{pid}_j^{s_j}$.

The security against active adversaries is only concerned about *fresh* sessions.

Definition 1. An instance Π_i^s that participated in a key establishment with users $\text{pid}_i^s = \{U_{u_1}, \dots, U_{u_r}\} \ni U$ is referred to as fresh if none of the following is true:

- For any $U_j \in \{U_{u_1}, \dots, U_{u_r}\}$ a query $\text{Corrupt}(U_j)$ was executed before a query of the form $\text{Send}(U_\ell, *, *)$ has taken place, with $U_\ell \in \{U_{u_1}, \dots, U_{u_r}\}$.
- The adversary has queried $\text{Reveal}(U_i, s)$ or $\text{Reveal}(U_j, t)$ where Π_i^s and Π_j^t are partnered.

The advantage Adv_A of an adversary A in attacking a key establishment protocol P is now defined as $\text{Adv}_A := |2 \cdot \text{Succ} - 1|$, where Succ is the probability of success A has on guessing the choice of the Test oracle, queried on (U, i) such that Π_i^s is fresh.

Security Against Malicious Participants. Three additional properties for a key agreement protocol to resist insider attacks are called integrity, entity authentication and key contribution. We will concentrate on the first two properties, as key contribution can only be achieved by a transformation of the key computation that we do not modify in our framework. A protocol that provides key contribution is usually called *key agreement*.

Definition 2. A group key establishment protocol fulfills integrity if all instances that accept with the same session identifier sid with overwhelming probability compute the same session keys sk , and hold a pid -value that includes at least all uncorrupted users that have accepted with session identifier sid .

Definition 3. To an instance $\Pi_i^{s_i}$ strong entity authentication is provided if with overwhelming probability both, $\text{acc}_i^{s_i} = \text{true}$ and for all uncorrupted $U_j \in \text{pid}_i^{s_i}$ exists an oracle $\Pi_j^{s_j}$ with $\text{sid}_j^{s_j} = \text{sid}_i^{s_i}$ and $U_i \in \text{pid}_j^{s_j}$.

3 Tseng’s Protocol – Burmester-Desmedt Secure Against Malicious Participants

Tseng [19] presents a variant of the Burmester-Desmedt ring protocol [6], that, if an authenticated broadcast channel is given, allows to detect and even to identify malicious participants. The system parameters are a cyclic group $\mathbf{G} = \langle g \rangle$ of order q with a hard dlog problem. The session key sk will be an element of \mathbf{G} . The protocol can then be summarized as follows: In the first round each participant U_i selects a random $x_i \in \mathbb{Z}_q$ and broadcasts $y_i = g^{x_i}$. In the second round U_i waits for all y_j , computes $z_i = (y_{i+1}/y_{i-1})^{x_i}$ and a non-interactive (zero-knowledge) proof that indeed (y_{i+1}/y_{i-1}) was raised to the power x_i . Each U_i broadcasts z_i and the proof. Finally, the messages and proofs are checked and the key is computed by participant U_i as in Burmester-Desmedt’s protocol as $\text{sk}_i = (y_{i-1})^{n x_i} \cdot z_i^{n-1} \cdot z_{i+1}^{n-2} \cdot \dots \cdot z_{i-2}$. A more detailed overview is given in Figure 1.

This protocol is proven secure against passive adversaries and malicious insiders under the assumption of an authenticated broadcast channel. The protocol

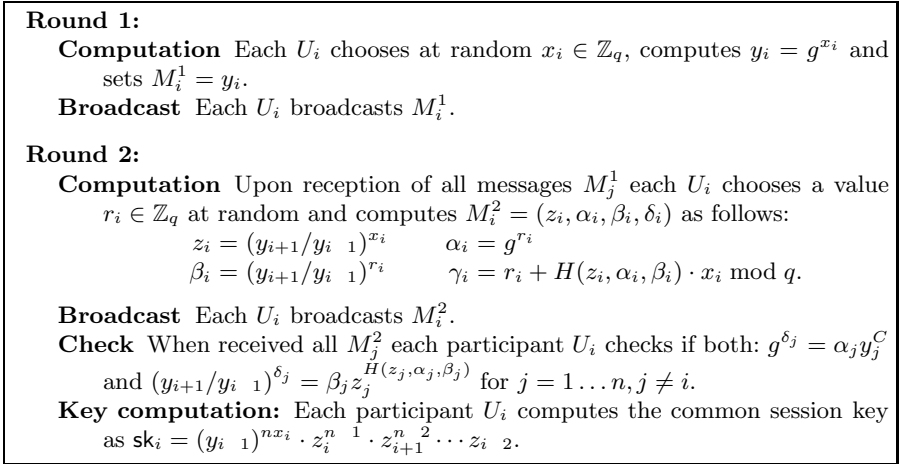


Fig. 1. The protocol of Tseng [19]

is not secure against active adversaries in the models [4, 5] or in [14]. Such an active adversary is allowed to selectively block or replace individual messages, and similarly, a malicious insider can send different messages to different participants instead of broadcasting. To “broadcast” in the protocol description in [14] stands merely for addressing the message to all other protocol participants.

We show that Tseng’s protocol is not secure against malicious participants in an unauthenticated network, thus, if using the compiler of [14] the protocol gets secure against active adversaries but loses security against malicious participants because of the following attack:

An Attack with Malicious Participants. The attack is similar to the attack in [19] to Burmester and Desmedt’s protocol, but has to be prepared more carefully. For simplicity, we assume a protocol run between 4 participants U_1, \dots, U_4 . Let now U_1 be corrupted. In Round 1, U_1 will, beside the random value x_1 , choose an additional random value \tilde{x}_1 and send $\tilde{y}_1 = g^{\tilde{x}_1}$ to U_3 , while sending honestly $y_1 = g^{x_1}$ to U_2 and U_4 . Since U_2 and U_4 are the only participants that use y_1 in Round 2 this will not influence the messages sent in Round 2 by U_2, U_3 or U_4 . By this, U_1 can in Round 2 send the value $\tilde{z}_1 = (y_2/y_4)^{\tilde{x}_1}$ and a valid proof to U_3 and send $z_1 = (y_2/y_4)^{x_1}$ with a valid proof to U_2 and U_4 . Thus, U_3 will compute a different key than U_2 and U_4 , because he uses the wrong value \tilde{z}_1 . Depending on if the session identifiers are equal or not this violates the key integrity or strong entity authentication.

The protocol can be transformed in a protocol that is secure against active adversaries and malicious insiders by applying both protocol compilers from [14, 13]. However, the protocol will then turn out to be a 4 round protocol as both compilers add one round to the protocol. We will introduce a framework that maintains the number of rounds, yet achieves security against malicious insiders and active adversaries.

4 The Framework

4.1 Session Identifier

Our framework is shown in Figure 2. We will now motivate the session identifier construction by studying its role with regard to the security against malicious participants. While proving security against passive adversaries for stateless protocols requires only one instance of the protocol – further rounds can be simulated by the adversary himself – for considering security against active adversaries, multiple instances of the participants are needed. To define which instances finally are allowed to know the established session identifier, the models [1, 5, 7, 18] introduce the *session identifier*. The session identifiers play a key role for the security proof, as they define when the adversary may know the session key for a trivial reason and when it constitutes a successful attack. Unfortunately, the session identifier often did not get the needed attention, what eventually invalidates the given proof as shown recently in [9] and [10]. We will continue the analysis of the session identifier with focus on malicious participants.

The construction of the session identifier varies between the proposals. Two common suggestions are a *pre-agreed session identifier*, and a *session identifier build from the concatenation of all messages* sent by the participants.

Both methods have advantages for security against malicious participants. The pre-agreed session identifier gives *entity authentication* almost for free: The session identifier can be included in authenticated messages and thereby every participant is convinced that his partners take part in a session with the same

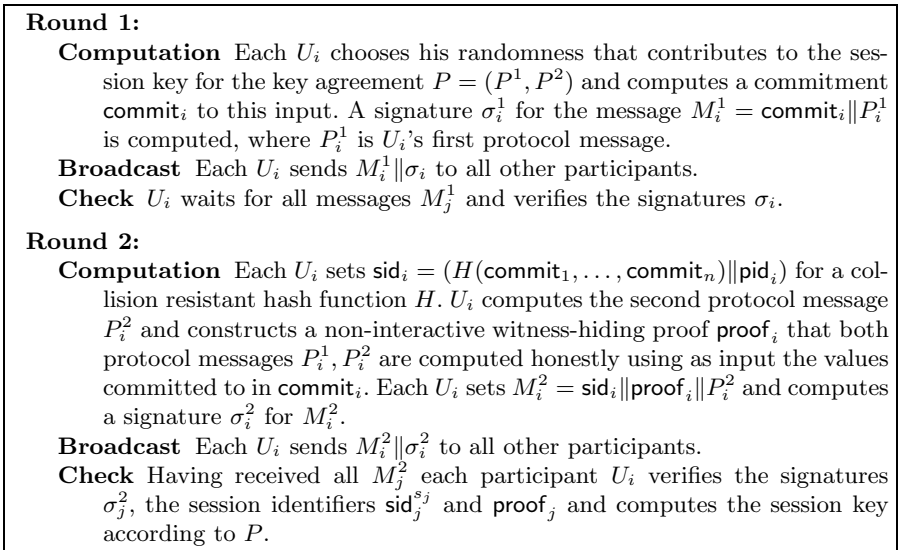


Fig. 2. A framework for malicious-participant secure AGKA

session identifier. In contrast, this is not possible if the session identifier is build up from a concatenation of the messages, as in this case the session identifier is defined only after the last message was sent. So, it can obviously not be included in a message of the protocol.

In the case of a session identifier build up from all protocol messages the session identifier serves as a commitment to the established session key. A single transcript of protocol messages can only produce one session key. Thus, all participants that accept a session with a matching session identifier can be sure that they accepted the same key – in this case *integrity* is obtained almost for free. In contrast, a pre-agreed session identifier obviously cannot be a commitment to the session key.

The following construction of the session identifier unifies the advantages of both versions:

- In the first round all participants commit to their randomness they use in the protocol which will influence the session key.
- The session identifier is build up from a concatenation of all commitments, or alternatively the result of a collision resistant hashfunction H applied to all commitments.
- In the following round the session identifier is included in an authenticated message.

In this way, the session identifier is known to all participants before the end of the protocol and can be used for entity authentication — on the other hand, the session identifier serves as a commitment to the session key.

Lemma 1. *With this method, if U_i chooses his inputs uniformly at random from a super-polynomial set, no two instances of a user U_i will with overwhelming probability compute equal session identifiers.*

Proof. The session identifier depends solely on the messages of the first round. Given an instance Π_i^s of user U_i , this instance will choose its input uniformly at random, compute a commitment commit_i^s and apply a collision resistant hashfunction H to all commitments to obtain the session identifier.

If an instance Π_i^t computes the same session identifier using its own commitment commit_i^t , then one of the following must be true:

- a collision of H occurs
- $\text{commit}_i^s = \text{commit}_i^t$ and $\text{input}_i^s \neq \text{input}_i^t$
- $\text{input}_i^s = \text{input}_i^t$

By the collision freeness of the hash function, the binding property of the commitment and the super-polynomial size of the set of possible inputs the probability of those events is negligible. Thus, the statement is true with overwhelming probability. \square

4.2 Embedding Protocols into the Framework

Given a key agreement protocol P consisting of 2 rounds P^1, P^2 that is secure against passive adversaries, we describe a framework to obtain a protocol that is

secure against active adversaries and malicious participants and still needs only 2 rounds. To achieve security against malicious participants, the session identifier for this protocol can be constructed as described above. All protocol participants have to commit on the random choices they will use for the protocol in Round 1 with a (non-interactive) commitment scheme that is computationally hiding and binding. Moreover, they have to prove in Round 2 that they behave honestly with respect to the commitments. In general this can be done with a non-interactive witness-hiding proof as we assumed in our framework. Though, in many cases it might be secure to reveal the commitments to protocol participants. All messages in the protocol will be signed with a signature scheme that is existentially unforgeable under chosen message attacks.

4.3 Proof of Security

Given a 2 round key agreement P that is secure against passive adversaries, we apply the framework from Figure 2 to P and obtain the protocol P_F . By *key agreement* we assume that every participant chooses his contribution to the key from a super-polynomial sized set and no collusion of participants (other than all participants together) can enforce a certain session key with non-negligible probability. Furthermore, we assume that P is stateless such that all protocol executions are independent.

The following lemma expresses basically that the adversary cannot successfully forge or replay messages in the name of an uncorrupted user.

Lemma 2. *If U_i is uncorrupted, then with overwhelming probability no instance $\Pi_j^{s_j}$ with $U_i \in \text{pid}_j^{s_j}$ will accept in an execution of protocol P_F , if not both messages appearing to be from U_i were sent by a single instance $\Pi_i^{s_i}$.*

Proof. If not both messages were sent by the same instance $\Pi_i^{s_i}$ this session is only accepted if one of the following cases happened:

1. One message was sent from another of U_i 's instances (possibly a replay).
2. A message was not generated by any of U_i 's instances.

For being successful in the first case, the session identifier of this session has to match the session of a previous session. However, this can be excluded with overwhelming probability from Lemma 1. The latter case only occurs with a negligible probability, because both messages are signed, and so, the forgery of a message includes the forgery of the signature scheme. However, the signature scheme is existentially unforgeable by assumption. \square

With the help of this lemma we can prove the security of the protocol P_F against malicious participants and active adversaries stipulated in the following propositions.

Proposition 1. *The protocol P_F provides entity authentication according to Definition 3 and integrity according to Definition 2.*

Proof. Entity authentication. Let Π_i^s be an arbitrary instance of an uncorrupted participant U_i that has accepted a session of the protocol P_F with session identifier sid_i . Let $U_j \in \text{pid}_i$ be any uncorrupted participant that Π_i^s believes took part in this session. By Lemma 2, Π_i^s must have received two messages of U_j with overwhelming probability and so an instance of U_j with $\text{sid}_j = \text{sid}_i$ implying $\text{pid}_j = \text{pid}_i$ must exist with overwhelming probability.

Integrity. Given two instances $\Pi_i^{s_i}$ and $\Pi_j^{s_j}$ of uncorrupted participants U_i, U_j who have accepted a session with the corresponding session identifier sid . By Lemma 2, in Round 2 of the protocol with overwhelming probability they have received signed messages of each other. Because this message includes the session identifier sid , the session identifiers $\text{sid}_i^{s_i} = \text{sid}_j^{s_j} = \text{sid}$ must indeed be identical. By construction of sid , both participants agree on the set pid and, unless a collision of the hash function H occurred, know the same values commit_i for every $U_i \in \text{pid}$. Both can check the proofs that all values needed for the key computation in the protocol messages were with overwhelming probability honestly computed with respect to $\text{commit}_1, \dots, \text{commit}_n$. Thus, they get consistent values for the key computation function and compute the same session key sk with overwhelming probability. \square

Proposition 2. *The protocol P_F is secure against active adversaries.*

Proof. Let $\Pi_i^{s_i}$ be the instance that is given in the call to the Test-Oracle. This oracle can have accepted by an Execute query or a sequence of Send queries.

In the first case the active adversary has no advantage, because already the passive adversary had access to the Execute oracle and all other sessions are independent so additional Send queries provide no information about the Test-session. By Lemma 2 it is also impossible to replay messages of the test session so that the receiver accepts. The adversary may combine the instances to a session at will without being able to break the scheme as long as the security of the scheme is independent of the number of participants or the set of users is of constant size.

We consider the case that the instance $\Pi_i^{s_i}$ accepted after a sequence of Send queries. The adversary has two possibilities to win the game

1. The adversary has information about $\text{sk}_i^{s_i}$ by sending forged messages to $\Pi_i^{s_i}$,
2. The adversary has information about $\text{sk}_i^{s_i}$ by revealing the session key $\text{sk}_j^{s_j}$ of a non-partnered oracle $\Pi_j^{s_j}$. To be related with sk_i the adversary must send forged messages to $\Pi_j^{s_j}$, again because different execution are independent.

In both cases the adversary has to query Test respectively. Reveal to an instance that has accepted. To be able to get additional information compared to a passive adversary, the adversary has to address that oracle with a manipulated message. Instance $\Pi_i^{s_i}$ will only be fresh if $\Pi_i^{s_i}$ accepted and no participant $U_j \in \text{pid}_i$ was addressed with a $\text{Send}(U_j, \cdot)$ query after a $\text{Corrupt}(U_k)$ query for an oracle $U_k \in \text{pid}_i$. Thus, all alleged senders of messages to instances that are partnered with $\Pi_i^{s_i}$ are uncorrupted at the time the message arrives. However, by Lemma 2, it is only possible with negligible probability to cheat in this case. \square

5 Applications of the Framework

The construction of the proof in Round 2 will not be efficient in general, however, for certain protocols a flexible application of the framework allows the construction of efficient and secure protocols.

In the protocol of Tseng (Figure 1), the first protocol message $M_i^1 = y_i = g^{x_i}$ represents already a commitment to the random choice x_i , and the proof given in Round 2 in Tseng's protocol guarantees the honest construction of the message M_i^2 as already proven in [19]. Enriched by the developed session identifier construction, this protocol fits into the framework – it is actually an application of the framework to the protocol of Burmester and Desmedt. The result is a new two round protocol secure against malicious insiders and active adversaries in a point-to-point network. This protocol allows additionally to identify the cheater, because the commitments and proofs cover all random inputs of the participants.

Another example is the secure protocol of Bohli et al. in [3]. Every participant U_i contributes a random value x_i and the session key is computed as $sk = H(x_1, \dots, x_n)$ (H assumed to behave like a random oracle). All but one participant broadcast their input x_i unencrypted in Round 1, only one distinguished participant U_n sends a commitment to his input x_n in Round 1 and distributes x_n encrypted among the participants in Round 2. The participants that broadcast their input in the first round are by this already committed and we can set $commit_i = x_i$ for $1 \leq i \leq n - 1$ and only, as in the protocol, $commit_n = H(x_n)$. The session identifier can be built from the commitments as proposed in the framework. Because also the input of the distinguished participant U_n is subsequently revealed to all protocol participants, there is no need for a proof – every participant can afterwards check the correctness of the commitment $commit_n$.

6 Conclusions

The protocol of Tseng was shown to be insecure against malicious participants in a point-to-point network. Nevertheless, the study of the protocol and the session identifier in group key agreement protocols led us to a framework for the construction of protocols that are secure against malicious participants and active adversaries. In contrast to existing protocol compilers, our framework does not increase the protocol's number of rounds. We could show that a flexible use of the framework allows the construction of efficient and secure protocols.

Acknowledgments. The author wishes to thank Rainer Steinwandt, Jörn Müller-Quade and the anonymous reviewers for their valuable comments.

References

1. M. Bellare, D. Pointcheval, and P. Rogaway. Authenticated Key Exchange Secure Against Dictionary Attacks. In *EUROCRYPT 2000*, volume 1807 of *LNCS*, pages 139–155. Springer, 2000.

2. M. Bellare and P. Rogaway. Entity Authentication and Key Distribution. In *CRYPTO '93*, volume 773 of *LNCS*, pages 232–249. Springer, 1993.
3. J.-M. Bohli, M. I. González Vasco, and R. Steinwandt. Secure group key establishment revisited. Cryptology ePrint Archive, Report 2005/395, 2005. <http://eprint.iacr.org/2005/395/>.
4. E. Bresson, O. Chevassut, and D. Pointcheval. Provably Authenticated Group Diffie-Hellman Key Exchange - The Dynamic Case. In *ASIACRYPT 2001*, volume 2248 of *LNCS*, pages 290–309. Springer, 2001.
5. E. Bresson, O. Chevassut, D. Pointcheval, and J.-J. Quisquater. Provably Authenticated Group Diffie-Hellman Key Exchange. In *ACM CCS*, pages 255–264. ACM Press, 2001.
6. M. Burmester and Y. Desmedt. A Secure and Efficient Conference Key Distribution System. In *EUROCRYPT '94*, volume 950 of *LNCS*, pages 275–286. Springer, 1995.
7. R. Canetti and H. Krawczyk. Analysis of Key-Exchange Protocols and Their Use for Building Secure Channels. In *EUROCRYPT 2001*, volume 2045 of *LNCS*, pages 453–474. Springer, 2001.
8. K.-K. R. Choo, C. Boyd, and Y. Hitchcock. Examining Indistinguishability-Based Proof Models for Key Establishment Protocols. In *ASIACRYPT 2005*, volume 3788 of *LNCS*, pages 585–604. Springer, 2005.
9. K.-K. R. Choo, C. Boyd, Y. Hitchcock, and G. Maitland. On Session Identifiers in Provably Secure Protocols. In *SCN 2004*, volume 3352 of *LNCS*, pages 351–366. Springer, 2005.
10. K.-K. R. Choo and Y. Hitchcock. Security Requirements for Key Establishment Proof Models. In *ACISP 2005*, volume 3574 of *LNCS*, pages 429–442. Springer, 2005.
11. D. Gollmann. Insider Fraud (Position Paper). In *Security Protocols, 6th International Workshop*, volume 1550 of *LNCS*, pages 213–219. Springer, 1998.
12. I. Ingemarsson, D. T. Tang, and C. K. Wong. A Conference Key Distribution System. *IEEE Transactions on Information Theory*, 28(5):714–720, 1982.
13. J. Katz and J. S. Shin. Modeling Insider Attacks on Group Key-Exchange Protocols. In *ACM CCS*, 2005.
14. J. Katz and M. Yung. Scalable Protocols for Authenticated Group Key Exchange. In *CRYPTO 2003*, volume 2729 of *LNCS*, pages 110–125. Springer, 2003.
15. B. Klein, M. Otten, and Th. Beth. Conference Key Distribution Protocols in Distributed Systems. In *Cryptography and Coding IV*, pages 225–241. IMA, 1993.
16. C.-H. Li and J. Pieprzyk. Conference Key Agreement from Secret Sharing. In *ACISP'99*, volume 1587 of *LNCS*, pages 64–76. Springer, 1999.
17. V. Shoup. On Formal Models for Secure Key Exchange. Cryptology ePrint Archive, 1999. <http://eprint.iacr.org/1999/012>.
18. M. Steiner. *Secure Group Key Agreement*. PhD thesis, Universität des Saarlandes, 2002. http://www.semper.org/sirene/publ/Stein_02.thesis-final.pdf.
19. Y.-M. Tseng. A Robust Multi-Party Key Agreement Protocol Resistant to Malicious Participants. *The Computer Journal*, 48(4):480–487, 2005.
20. W.-G. Tzeng. A Practical and Secure Fault-Tolerant Conference-Key Agreement Protocol. In *PKC 2000*, volume 1751 of *LNCS*, pages 1–13. Springer, 2000.
21. W.-G. Tzeng and Z.-J. Tzeng. Round-Efficient Conference Key Agreement Protocols with Provable Security. In *ASIACRYPT 2000*, volume 1976 of *LNCS*, pages 614–627. Springer, 2000.

BGN Authentication and Its Extension to Convey Message Commitments

Yuen-Yan Chan¹ and Jin Li^{1,2}

¹ Department of Information Engineering,
Chinese University of Hong Kong,
Shatin, N.T., Hong Kong
{yychan, jinli}@ie.cuhk.edu.hk

² School of Mathematics and Computational Science,
Sun Yat-Sen University, Guangzhou, 510275, P.R. China
sysjinli@yahoo.com.cn

Abstract. We instantiate the cryptosystem proposed by Boneh, Goh, and Nissim in TCC'05 [5] into an entity authentication scheme, in which an entity is authenticated by an interactive zero-knowledge proof on its private key. Completeness and soundness of our scheme is supported by the indistinguishability of BGN ciphertexts of sums and products, which essentially relies on the semantic security of the BGN cryptosystem. We further extend our scheme so that the authentication conveys Pedersen commitments on a message, while the BGN authentication serves the ‘proving you know how to open’ functionality for the commitment. Our message commitment scheme is both statistically hiding and computationally binding provided the subgroup decision problem is hard.

1 Introduction

Entity¹ authentication is a process whereby a verifier is assured of the identity of a prover after an interactive acquisition of corroborative evidence. There can be two major approaches for entity authentication: cryptographic (such as the use of standard passwords and digital certificates) and non-cryptographic (such as the use of biometrics techniques). Cryptographic protocols for entity authentication can further be classified as weak, strong, and zero-knowledge based [2], which corresponds to the identification of an entity by means of passwords (such as [18, 20, 17]), challenge-response protocols (by symmetric-key techniques, such as [19, 24, 28], and by public-key techniques, such as [6, 25], and interactive zero-knowledge proofs (such as [10, 15, 14, 23, 27])).

In this paper, we propose an authentication scheme that based on an interactive zero-knowledge proof on the prover’s private key generated in the BGN cryptosystem settings. In our scheme the challenge is a pair of ciphertexts of

¹ Acknowledgement to Hong Kong Research Grant Council’s Earmarked Grants 4232-03E and 4328-02E for sponsorship. Part of the work of the second co-author was done while visiting The Chinese University of Hong Kong.

two messages encrypted by the prover's public key. Instead of asking the prover to decrypt the challenge and return the corresponding plaintext (a technique widely used in the challenge-response based authentication protocols, such as [6]), our scheme requires the prover to distinguish between the ciphertexts of the sum of and that of the product of the pair of corresponding plaintexts in the challenge. In this way, the plaintext messages can be committed and hidden, and be opened after an arbitrary time interval. In particular, we extend our scheme to convey committed messages such that it becomes a commitment scheme, a method by which the prover can *commit* to a message to the verifier without revealing the message, and the commitment can later be *opened* by the prover and verified by the verifier. Furthermore, we utilize the proposed authentication scheme to provide the '*prove-you-know-how-to-open*' functionality, which is an add-on functionality to the message commitment scheme proposed in [12].

1.1 Related Works

Entity Authentication Schemes by Prove of Knowledge. Fiat and Shamir first introduced the use of interactive proof for entity authentication [15]. Later, the notion of interactive proof and zero-knowledge has been formalized by Goldwasser *et. al.* [16]. Schnorr further applied interactive zero-knowledge proof in identification and signatures in smartcards [27]. Some other schemes were also proposed since then, such as [10, 23]. It is worth notice that Canetti *et. al.* raised the *reset attack* [8], an attack that breaks the security of Fiat-Shamir like identification schemes [15, 14, 27] whenever the prover is resettable. Fortunately, the scheme proposed in this paper is not constructed with the Fiat-Shamir technique.

Message Commitment Schemes. Since the introduction of the classical *coin flipping by telephone* problem, which could be solved by bit commitment [4], a number of commitment protocols have been proposed, such as [3, 7, 13, 22, 26]. Message commitment schemes enables a prover to commit a value and give the commitment to the verifier, who cannot see the value until the prover opens it. Two basic requirements of commitment schemes are *hiding* and *binding*: that the committed message is hidden from the verifier while the prover is committed to the message. Later, various security properties are introduced to commitment schemes including *non-malleability* [13, 21], *universally composability* [11], and the *mercurial property* [9].

Previous Works in BGN Cryptosystem. Our result is based on the the cryptosystem newly proposed by Boneh, Goh, and Nissim in TCC'05 [5] which introduces the *Subgroup Decision Problem*, a new trapdoor structure that supports a public key encryption with dual homomorphic properties. As of the day of completion of this paper, three pieces of works that follow the results of Boneh *et. al.* but not yet formally unpublished are available. Groth *et. al.* presented a non-interactive zero knowledge which is secure under the universal composability framework [29]. Adida *et. al.* applied the Boneh-Goh-Nissim cryptosystem in mixnets and proposed the obfuscated ciphertext mixing [1]. Recently Wei presented a new assumption on the computational conjugate subgroup mem-

bers (CCSM) and proposed a new signature which security is reducible to the assumption [30].

1.2 Our Contribution

We define and prove the indistinguishability of BGN ciphertexts of sums and products and use the result to construct an interactive zero-knowledge proof-based authentication scheme (BGNAuth). We further extend our scheme to obtain a message commitment scheme (BGNMC), with which BGNAuth can be used in conjunction to provide the ‘proving-you-know-how-to-open’ functionality for the commitment. We also prove that our scheme is both statistically hiding and computationally binding provided the subgroup decision problem is hard.

2 Preliminaries

2.1 Notations

Following the notations in [5], define a bilinear group \mathbb{G} with the generator g as follow: let \mathbb{G} and \mathbb{G}_1 be two multiplicative cyclic groups of order n where $n = q_1q_2 \in \mathbb{Z}$ and q_1, q_2 are some τ -bit primes for some security parameter $\tau \in \mathbb{Z}^+$, and $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_1$ is a bilinear map and $e(g, g)$ is a generator of \mathbb{G}_1 .

2.2 Subgroup Decision Problem

The *Subgroup Decision Problem* is the one defined below. Let \mathcal{G} be an algorithm and $\mathcal{G}(\tau)$ outputs the tuple $(q_1, q_2, \mathbb{G}, \mathbb{G}_1, e)$ where $q_1, q_2, \mathbb{G}, \mathbb{G}_1$ and e are defined above. Given $(n, \mathbb{G}, \mathbb{G}_1, e)$ and $x \in \mathbb{G}$, output ‘1’ if order of x is q_1 and output ‘0’ otherwise. For an adversary \mathcal{A} , define $\text{Adv}_{\mathcal{A}}^{\text{sd}}(\tau)$, the advantage of \mathcal{A} in solving the subgroup decision problem as

$$\left| \Pr[\mathcal{A}(n, \mathbb{G}, \mathbb{G}_1, e, x) = 1 : (q_1, q_2, \mathbb{G}, \mathbb{G}_1, e) \leftarrow \mathcal{G}(\tau), n = q_1q_2, x \leftarrow \mathbb{G}] - \Pr[\mathcal{A}(n, \mathbb{G}, \mathbb{G}_1, e, x^{q_2}) = 1 : (q_1, q_2, \mathbb{G}, \mathbb{G}_1, e) \leftarrow \mathcal{G}(\tau), n = q_1q_2, x \leftarrow \mathbb{G}] \right|.$$

Definition 1 (Subgroup Decision Assumption). \mathcal{G} satisfies the subgroup decision assumption if for any polynomial time adversary \mathcal{A} we have $\text{Adv}_{\mathcal{A}}^{\text{sd}}(\tau)$ being a negligible function in τ .

2.3 Boneh-Goh-Nissim Cryptosystem

The BGN cryptosystem is a tuple of algorithms (KeyGen, Encrypt, Decrypt) where:

- KeyGen(τ): Run $\mathcal{G}(\tau)$ for a given security parameter $\tau \in \mathbb{Z}^+$ to generate $(q_1, q_2, \mathbb{G}, \mathbb{G}_1, e)$. Let $n = q_1q_2$. Pick generators $g, u \stackrel{R}{\leftarrow} \mathbb{G}$ and let $h = u^{q_2}$. Output public key $\mathcal{PK} = (n, \mathbb{G}, \mathbb{G}_1, e, g, h)$ and private key $\mathcal{SK} = q_1$.

- $\text{Encrypt}(\mathcal{PK}, m)$: To encrypt a message $m \in \mathbb{Z}_T$ where $T < q_2$ with public key \mathcal{PK} , pick a random $r \in \mathbb{Z}_n$ and compute $C = g^m h^r \in \mathbb{G}$. Output C as the ciphertext.
- $\text{Decrypt}(\mathcal{SK}, C)$: To decrypt a message C using the private key \mathcal{SK} , compute $C^{q_1} = (g^m h^r)^{q_1} = (g^{q_1})^m$ and recover m by computing the discrete log of C^{q_1} base g^{q_1} . This expects $\tilde{O}(\sqrt{T})$ with the Pollard’s lambda method and is efficient for small finite T .

Definition 2 (BGN Ciphertexts). A BGN ciphertext C of a message $m \in \mathbb{Z}_T$ is an output of $\text{Encrypt}(\mathcal{PK}, m)$ with \mathcal{PK} generated from $\text{KeyGen}(\tau)$ for some security parameter τ .

2.4 Dual Homomorphic Property of BGN Cryptosystem

Given two BGN ciphertexts $C_1, C_2 \in \mathbb{G}_1$ of messages $m_1, m_2 \in \mathbb{Z}_T$. Set $g_1 = e(g, g)$ and $h_1 = e(g, h)$ where g_1 has order n and h_1 has order q_2 . Also write $h = g^{\alpha q_2}$ for some $\alpha \in \mathbb{Z}$. For random $r, r_1, r_2 \in \mathbb{Z}_n$, define operators \oplus and \odot that work as follow:

$$C_1 \oplus C_2 = C_1 C_2 h^r \in \mathbb{G}$$

$$C_1 \odot C_2 = e(C_1, C_2) h_1^r = e(g^{m_1} h^{r_1}, g^{m_2} h^{r_2}) h_1^r = g_1^{m_1 m_2} h_1^{\tilde{r}} \in \mathbb{G}_1$$

where $\tilde{r} = m_1 r_2 + m_2 r_1 + \alpha q_2 r_1 r_2 + r$ is distributed uniformly in \mathbb{Z}_n . We can see that results from $C_1 \oplus C_2$ and $C_1 \odot C_2$ are indeed the BGN ciphertexts of $m_1 + m_2$ and $m_1 m_2$ respectively. The above gives the dual homomorphic property (both additively homomorphic and multiplicative homomorphic) of BGN cryptosystems. From now on we denote $C_S = C_1 \oplus C_2$ and $C_P = C_1 \odot C_2$ for two messages $m_1, m_2 \in \mathbb{Z}_T$.

2.5 Security of BGN Cryptosystem

Theorem 1 (Semantic Security of BGN Cryptosystem). The BGN cryptosystem is semantically secure if \mathcal{G} satisfies the subgroup decision assumption.

Proof. Given in [5] Section 3.2.

Consider the BGN cryptosystem. For two inputs m_1, m_2 and the corresponding BGN ciphertexts $C_1, C_2 \in \mathbb{G}$, let $C_S, C_P \in \mathbb{G}_1$ as defined in Section 2.4. For $C \in \{C_S, C_P\}$, consider the following distinguisher:

Definition 3 (Distinguisher for (C_S, C_P)). A distinguisher \mathcal{D} for (C_S, C_P) is a probabilistic polynomial time algorithm that takes $C \in \{C_S, C_P\}$ as input and outputs (i) $\mathcal{D}(C, C_S) = 1$ iff $C = C_S$; (ii) and $\mathcal{D}(C, C_P) = 1$ iff $C = C_P$. We say \mathcal{D} distinguishes C_S and C_P with advantage $\text{Adv}_{\mathcal{D}} > 0$ if

$$\text{Adv}_{\mathcal{D}} = |\text{Prob}[\mathcal{D}(C, C_S) = 1] - \text{Prob}[\mathcal{D}(C, C_P) = 1]|.$$

Definition 4 (Indistinguishability of C_S and C_P). For C_S, C_P defined above, C_S and C_P are said to be polynomially indistinguishable if there exist no distinguisher for (C_S, C_P) with non-negligible advantage $\text{Adv}_{\mathcal{D}} > 0$.

We have the following theorem for C_S, C_P as defined above. We defer its proof as a security result in Section 5.

Theorem 2 (Indistinguishability of BGN Ciphertexts of Sums and Products). The BGN ciphertexts C_S, C_P of messages $m_1, m_2 \in \mathbb{Z}_T$ are polynomially indistinguishable provided the semantic security of BGN cryptosystem.

3 Security Model

3.1 Syntax

BGN Authentication (BGNAuth). On top of the the BGN cryptosystem (KeyGen, Encrypt, Decrypt), we define the syntax for *BGN Authentication (BGNAuth)* as the tuple (Setup, Witness, Challenge, Response, Verify) where:

- **Setup**(τ) $\mapsto (\mathcal{PK}, \mathcal{SK})$: For a security parameter τ , runs KeyGen to obtain $\mathcal{PK} = (n, \mathbb{G}, \mathbb{G}_1, e, g, h)$ and $\mathcal{SK} = q_1$.
- **Witness**(\mathcal{PK}, m_1, m_2) $\mapsto (C_1, C_2)$: For the BGN public key \mathcal{PK} , and $m_1, m_2 \in \mathbb{Z}_T \setminus \{m_1 + m_2 = m_1 m_2 \text{ mod } T\}$, compute $C_1 = \text{Encrypt}(\mathcal{PK}, m_1)$ and $C_2 = \text{Encrypt}(\mathcal{PK}, m_2)$. Output C_1, C_2 .
- **Challenge**($\mathcal{PK}, C_1, C_2, b$) $\mapsto C$: For a random $b \in \{0, 1\}$, the BGN public key \mathcal{PK} and the BGN ciphertexts C_1, C_2 , set $C = C_1 \oplus C_2$ if $b = 0$ and set $C = C_1 \odot C_2$ otherwise. Output C .
- **Response**($\mathcal{SK}, m_1, m_2, C$) $\mapsto b' \in \{0, 1, \perp\}$: With the BGN private key \mathcal{SK} and the BGN ciphertext C , obtain $M = \text{Decrypt}(\mathcal{SK}, C)$. Output $b' = 0$ if $M = m_1 + m_2 \text{ mod } T$, output $b' = 1$ if $M = m_1 m_2 \text{ mod } T$, output $b' = \perp$ otherwise.
- **Verify**(b, b') $\mapsto \{0, 1\}$: Output 1 if $b = b'$, output 0 otherwise.

Extension to Conveys Message Commitments. BGNAuth offers Pedersen commitment [26] on a message $m \in \mathbb{Z}_T$. Under this context, we define the following commitment scheme, called *BGN Message Commitment (BGNMC)*, by extending BGNAuth. BGNMC is a tuple (Setup, Commit, Open) where:

- **Setup**(τ) $\mapsto (\mathcal{PK}, \mathcal{SK})$: For a security parameter τ , runs KeyGen to obtain $\mathcal{PK} = (n, \mathbb{G}, \mathbb{G}_1, e, g, h)$ and $\mathcal{SK} = q_1$.
- **Commit**(\mathcal{PK}, m_1, m_2) $\mapsto (c_1, c_2, r_1, r_2)$: Same as Witness in BGNAuth. Except two random numbers $r_1, r_2 \in \mathbb{Z}_n$ used in the BGN encryption are at least τ -bit long and are stored for later use. $m_1 = m$ and m_2 can be any dummy message $\in \mathbb{Z}_T$ as long as $m + m_2 \neq m m_2 \text{ mod } T$. The outputs are $c_1 = g^{m_1} h^{r_1}$ and $c_2 = g^{m_2} h^{r_2} \in \mathbb{G}$.
- **Open**($\mathcal{PK}, m_1, m_2, r_1, r_2$) $\mapsto \{0, 1\}$: Computes $\tilde{c}_1 = g^{m_1} h^{r_1}$ and $\tilde{c}_2 = g^{m_2} h^{r_2}$ with $h \in \mathcal{PK}$. Output ‘1’ if $\tilde{c} = c$, output ‘0’ otherwise.

3.2 Security Notions

We have the following definitions on **completeness**, **soundness**, and **zero-knowledge** for an interactive zero-knowledge-ness proof-based authentication scheme.

Definition 5 (Completeness Property). *With given (τ, m_1, m_2) , honest V and honest P , completeness probability $Pr_{P,V}^{comp}$ is defined as:*

$$Pr[\text{Verify}(b', b) = 1 | (\mathcal{PK}, \mathcal{SK}) \leftarrow \text{SetUp}(\tau), (C_1, C_2) \leftarrow \text{Commit}(\mathcal{PK}, m_1, m_2), \\ (b, C) \leftarrow \text{Challenge}(\mathcal{PK}, C_1, C_2), b' \leftarrow \text{Response}(\mathcal{SK}, m_1, m_2, C)].$$

An interactive proof-based authentication scheme is complete if the authentication protocol succeeds with an overwhelming $Pr_{P,V}^{comp}$.

Definition 6 (Soundness Property). *With given (τ, m_1, m_2) , V , dishonest P^* , and a polynomial time extractor \mathcal{M} who extracts the private key \mathcal{SK}' with given \mathcal{PK} for P^* . The soundness probability $Pr_{P^*,V}^{sound}$ is defined as:*

$$Pr[\text{Verify}(b', b) = 1 | (\mathcal{PK}, \tilde{\mathcal{SK}}') \leftarrow \mathcal{M}, (C_1, C_2) \leftarrow \text{Commit}(\mathcal{PK}, m_1, m_2), \\ (b, C) \leftarrow \text{Challenge}(\mathcal{PK}, C_1, C_2), b' \leftarrow \text{Response}(\tilde{\mathcal{SK}}', m_1, m_2, C)].$$

Define $Adv_{\mathcal{M}} = Pr_{P^,V}^{sound}$ as the advantage of the extractor \mathcal{M} . An interactive proof-based authentication scheme is sound if $Adv_{\mathcal{M}}$ is negligible for all polynomial time extractors.*

Definition 7 (Zero-Knowledge Property). *With given (τ, m_1, m_2) , P , V , and a polynomial-time simulator \mathcal{S} . Denote $\Phi_{P,V} = \{C_1, C_2, C_b, b'\}$ the set of transcripts produced by P and V in a executing the BGN authentication. Denote $\tilde{\Phi}_{\mathcal{S},V} = \{\tilde{C}_1, \tilde{C}_2, \tilde{C}_b, b'\}$ the set of transcripts simulated by \mathcal{S} without interacting with P . An interactive proof-based authentication scheme is zero-knowledge if $\Phi_{P,V}$ and $\tilde{\Phi}_{\mathcal{S},V}$ are indistinguishable.*

We also define the **hiding** and **binding** properties for a message commitment scheme.

Definition 8 (Hiding Property). *For some public key \mathcal{PK} and uniform m, m' , a hiding message commitment scheme has indistinguishable commitments c, c' where $c = \text{Commit}(\mathcal{PK}, m)$ and $c' = \text{Commit}(\mathcal{PK}, m')$.*

Definition 9 (Binding Property). *For some public key \mathcal{PK} and uniform m, m' , a message commitment scheme is binding if it holds that for a running of at most polynomial-time, the probability that $c = \text{Commit}(\mathcal{PK}, m) = c' = \text{Commit}(\mathcal{PK}, m')$ and $c \neq c'$, is negligible.*

4 Constructions

4.1 BGN Authentication

BGNAuth is a standard ‘Witness-Challenge-Response’ zero-knowledge authentication protocol. Fig. 1 illustrates the protocol run between a prover P and a verifier V .

INPUT: $\text{SetUp}(\tau)$ is run beforehand with a security parameter τ to output keys $(\mathcal{PK}, \mathcal{SK})$. Both P and V gets \mathcal{PK} . V keeps \mathcal{SK} as her private key.

1. **WITNESS:** P performs the following:
 - (a) P selects $m_1, m_2 \in \{\mathbb{Z}_T\}$.
 - (b) P runs $\text{Commit}(\mathcal{PK}, m_1, m_2)$ to get C_1, C_2 .
 - (c) P sends C_1, C_2 to V and stores m_1, m_2 for later use.
2. **CHALLENGE:** V performs the following:
 - (a) V selects a random $b \in \{0, 1\}$.
 - (b) V runs $\text{Challenge}(\mathcal{PK}, C_1, C_2, b)$ to get C .
 - (c) V stores b for later use and sends C to V.
3. **RESPONSE:** P performs the following:
 - (a) P performs $\text{Response}(\mathcal{SK}, m_1, m_2, C)$ to get b .
 - (b) P sends b to V.

V authenticates P if $\text{Verify}(b, b) = 1$.

Remark: The protocol can be repeated for arbitrary number of times to convince V.

Fig. 1. BGN Authentication (BGNAuth)

4.2 Message Commitment Scheme Extended from BGN Authentication

BGNAuth conveys Pedersen commitments on two messages. The following protocol (BGN Message Commitment, BGNMC) illustrates the COMMIT and OPEN for the messages initiated in BGNAuth and is illustrated in Fig. 2. We also make a remark² on homomorphism and non-malleability of our scheme.

4.3 ‘Proving You Know How to Open’

In [12] Damgård *et. al.* proposed an associated protocol called ‘Proving You Know How to Open’ for their integer commitment scheme. Such protocol enables P to show that he can open a given commitment $c = g^x h^r$. In our scheme, BGNAuth (Fig. 1) can be run in conjunction with BGNMC (Fig. 2) so as to show that P can open the commitments.

5 Security Results

Proof (Theorem 2). Suppose a polynomial time distinguisher \mathcal{D} without \mathcal{SK} distinguishes C_S, C_P of some messages $x_1, x_2 \in \mathbb{Z}_T$ with advantage $\epsilon(\tau)$. We construct a polynomial time algorithm \mathcal{B} that breaks the semantic security of

² The commitments inherit dual homomorphism from BGN cryptosystem. This makes BGNMC malleable. So our scheme should be applied when homomorphism is desired over non-malleability, such as committing on overall price from the price of each item.

INPUT: $\text{SetUp}(\tau)$ is run beforehand with a security parameter τ to output keys $(\mathcal{PK}, \mathcal{SK})$. Both P and V gets \mathcal{PK} . V keeps \mathcal{SK} as her private key.

1. COMMIT: P performs the following:
 - (a) P selects $m_1, m_2 \in \{\mathbb{Z}_T\}$. In case only one message is to be committed (assume is m_1), selects dummy m_2 where $m_1 + m_2 \neq m_1 m_2 \pmod T$.
 - (b) P runs $\text{Commit}(\mathcal{PK}, m_1, m_2)$ to get c_1, c_2, r_1, r_2 .
 - (c) P sends c_1, c_2 to V and stores r_1, r_2 for later use.
2. OPEN: P and V perform the following:
 - (a) P gives m_1, m_2, r_1, r_2 to V.
 - (b) V accepts c_1, c_2 as commitments of m_1, m_2 respectively if $\text{Open}(\mathcal{PK}, m_1, m_2, r_1, r_2) = 1$.

Fig. 2. BGN Message Commitment (BGNMC)

the BGN cryptosystem with the same advantage. Algorithm \mathcal{B} works as follows: Simulator \mathcal{S} gives \mathcal{B} the public key $\mathcal{PK} = (n, \mathbb{G}, \mathbb{G}_1, e, g, h)$. \mathcal{B} generates two random values $x_1, x_2 \in \mathbb{Z}_T$, and outputs $m_0 = x_1 + x_2 \pmod T$ and $m_1 = x_1 \cdot x_2 \pmod T$ to \mathcal{S} . \mathcal{S} constructs $C_0 = \text{Encrypt}(\mathcal{PK}, m_0)$ and $C_1 = \text{Encrypt}(\mathcal{PK}, m_1)$. Note that $C_0 \in C_S$ and $C_1 \in C_P$ for $x_1, x_2 \in \mathbb{Z}_T$. \mathcal{S} selects a random $b \in \{0, 1\}$ and returns C_b to \mathcal{B} . \mathcal{B} gives \mathcal{PK}, x_1, x_2 , and C_b to \mathcal{D} . \mathcal{D} guess $b' \in \{0, 1\}$ and returns b' to \mathcal{B} . \mathcal{B} forward b' to \mathcal{S} . By the definition of \mathcal{D} who can distinguishes C_S, C_P with non-negligible advantage $\epsilon(\tau) > 0$, it follows that $\Pr[b = b'] > 1/2 + \epsilon(\tau)$. Hence \mathcal{B} breaks the semantic security of the BGN cryptosystem with advantage $\epsilon(\tau)$.

Theorem 3 (Completeness of BGN Authentication). *BGNAuth is complete.*

Proof. Obvious for honest V and honest P.

Theorem 4 (Soundness of BGN Authentication). *BGNAuth is sound.*

Proof. Let \mathcal{M} be a polynomial-time extractor with non-negligible $\text{Adv}_{\mathcal{M}}$. We can construct a polynomial-time algorithm \mathcal{B} that distinguishes BGN ciphertexts of sums and products (C_S, C_P) with the same advantage. Algorithm \mathcal{B} works as follows: Simulator \mathcal{S} gives \mathcal{B} the public key $\mathcal{PK} = (n, \mathbb{G}, \mathbb{G}_1, e, g, h)$. \mathcal{B} generates two random values $m_1, m_2 \in \mathbb{Z}_T$, and outputs C_1, C_2 , the corresponding BGN ciphertexts to \mathcal{S} . \mathcal{S} constructs $C'_0 = C_1 \oplus C_2$ and $C'_1 = C_1 \odot C_2$. \mathcal{S} select a random $b \in \{0, 1\}$ and returns C'_b to \mathcal{B} . \mathcal{B} gives \mathcal{PK}, m_1, m_2 , and C_b to \mathcal{M} . \mathcal{M} guess $b' \in \{0, 1\}$ for b and returns b' to \mathcal{B} . \mathcal{B} forward b' to \mathcal{S} . By the definition of \mathcal{M} who can extract \mathcal{SK}' from given \mathcal{PK} , it can produce $b' \leftarrow \text{Response}(\mathcal{SK}', m_1, m_2, C_b)$ with advantage $\text{Adv}_{\mathcal{M}}$ such that $\Pr_{\mathbb{P}^*_{\mathcal{P}, \mathcal{V}}}^{\text{sound}}$ is non-negligible. Hence \mathcal{B} distinguishes C_S, C_P with non-negligible advantage.

Theorem 5 (Zero-Knowledgeness of BGN Authentication). *BGNAuth is computationally zero-knowledge.*

Proof. (Sketch.) Follow from the fact that (C_1, C_2, C_b, b') , a valid set of transcripts obtained from the execution of BGNAuth by honest P and V, are computable in polynomial time from the public information \mathcal{PK} alone.

Theorem 6 (Hiding Property of BGN Message Commitment). *BGNMC is a hiding message commitment scheme.*

Proof. (Sketch.) Since the random number r used in Commit is at least τ -bit long where $\tau \geq \text{ord}(h)$, $c = \text{Commit}(\mathcal{PK}, m) = g^m h^r$ and $c' = \text{Commit}(\mathcal{PK}, m') = g^{m'} h^r$ are statistically close to uniform in $\langle h \rangle$ for any values of m, m' .

Theorem 7 (Binding Property of BGN Message Commitment). *BGNMC is a binding message commitment scheme.*

Proof. Suppose P^* can break the binding property and can create a commitment c with valid distinct openings (m, r) and (m', r') , that is $c = g^m h^r = g^{m'} h^{r'}$. Write $g = h^\alpha$ for some unknown $\alpha \in \mathbb{Z}$. Then we have $h^{\alpha m} h^r = h^{\alpha m'} h^{r'}$, that is $h^{\alpha(m-m') + (r-r')} = 1$. Hence $\text{ord}(h)$ must divide $M := \alpha(m - m') + (r - r')$ where M not necessarily equals n . We can now use P^* to break the Subgroup Decision Problem as follows: Given $(n, \mathbb{G}, \mathbb{G}_1, e)$ and $g \in \mathbb{G}$, P^* can compute M , a multiple of the order of $\langle h \rangle$ (the subgroup). That is, P^* has non-negligible advantage $\text{Adv}_{P^*}^{\text{sd}}(\tau)$.

6 Conclusion

In this paper, we have defined and proven the indistinguishability of BGN ciphertexts of sums and products, and constructed the BGN authentication scheme. We have also extended our scheme so that it conveys message commitments. Our scheme can be applied to situations that require authenticated commitments, such as e-bidding and e-voting. Commitments in our scheme are malleable but homomorphic. Therefore our scheme can be applied when homomorphism is desired over non-malleability, such as committing on a total price from the individual commitments on the price of each item, where additional authentication is in place when necessary.

References

1. Ben Adida and Douglas Wikstrom. Obfuscated ciphertext mixing. Cryptology ePrint Archive, Report 2005/394, November 2005. <http://eprint.iacr.org/>.
2. Scott A. Vanstone Alfred Menezes, Paul C. van Oorschot. *Handbook of Applied Cryptography*. CRC Press, 1996.
3. Donald Beaver. Adaptive zero knowledge and computational equivocation (extended abstract). In *STOC*, pages 629–638, 1996.
4. M. Blum. Coin flipping by telephone. *IEEE Spring COMPCOM*, pages 133–137, 1982.

5. D. Boneh, E.-J. Goh, and K. Nissim. Evaluating 2-dnf formulas on ciphertexts. In *Theory of Cryptography Conference, TCC*, pages 325–341, February 2005.
6. C. C. I. T. T. *Recommendation X.509. The Directory-Authentication*, 1988.
7. Ran Canetti and Marc Fischlin. Universally composable commitments. In *CRYPTO*, pages 19–40, 2001.
8. Ran Canetti, Oded Goldreich, Shafi Goldwasser, and Silvio Micali. Resettable zero-knowledge (extended abstract). In *STOC*, pages 235–244, 2000.
9. Melissa Chase, Alexander Healy, Anna Lysyanskaya, Tal Malkin, and Leonid Reyzin. Mercurial commitments with applications to zero-knowledge sets. In *EUROCRYPT*, pages 422–439, 2005.
10. Nicolas Courtois. Efficient zero-knowledge authentication based on a linear algebra problem minrank. In *ASIACRYPT*, pages 402–421, 2001.
11. I. Damgård and J. B. Nielsen. Perfect hiding and perfect binding universally composable commitment schemes with constant expansion factor. Technical report, BRICS Report Series RS-01-41, October 2001.
12. Ivan Damgård and Eiichiro Fujisaki. A statistically-hiding integer commitment scheme based on groups with hidden order. In *ASIACRYPT*, pages 125–142, December 2002.
13. Danny Dolev, Cynthia Dwork, and Moni Naor. Nonmalleable cryptography. *SIAM J. Comput.*, 30(2):391–437, 2000.
14. Uriel Feige, Amos Fiat, and Adi Shamir. Zero-knowledge proofs of identity. *J. Cryptology*, 1(2):77–94, 1988.
15. Amos Fiat and Adi Shamir. How to prove yourself: Practical solutions to identification and signature problems. In *CRYPTO*, pages 186–194, 1986.
16. Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof systems. *SIAM J. Comput.*, 18(1):186–208, 1989.
17. The Open Group. Unix. <http://www.unix.org/>.
18. N. M. Haller. The s/key one-time password system. In *Symposium on Network and Distributed System Security*, pages 151–157, 1994.
19. International Organization for Standardization. *ISO/IEC 9798-2*, July 1999.
20. Leslie Lamport. Password authentication with insecure communication. *Commun. ACM*, 24(11):770–772, 1981.
21. Moses Liskov, Anna Lysyanskaya, Silvio Micali, Leonid Reyzin, and Adam Smith. Mutually independent commitments. In *ASIACRYPT*, pages 385–401, 2001.
22. Moni Naor. Bit commitment using pseudorandomness. *J. Cryptology*, 4(2):151–158, 1991.
23. Moni Naor. Deniable ring authentication. In *CRYPTO*, pages 481–498, 2002.
24. Roger M. Needham and Michael D. Schroeder. Using encryption for authentication in large networks of computers. *Commun. ACM*, 21(12):993–999, 1978.
25. Roger M. Needham and Michael D. Schroeder. Authentication revisited. *Operating Systems Review*, 21(1):7, 1987.
26. Torben P. Pedersen. Non-interactive and information-theoretic secure verifiable secret sharing. In *CRYPTO*, pages 129–140, August 1991.
27. Claus-Peter Schnorr. Efficient identification and signatures for smart cards. In *CRYPTO*, pages 239–252, 1989.
28. J. G. Steiner, B. C. Neuman, and J. I. Schiller. Kerberos: An authentication service for open network systems. In *USENIX Winter*, pages 191–202, 1988.
29. J. Groth *et al.* Perfect non-interactive zero knowledge for np. Cryptology ePrint Archive, Report 2005/290, August 2005. <http://eprint.iacr.org/>.
30. Victor K. Wei. Signature from a new subgroup assumption. Cryptology ePrint Archive, Report 2005/429, November 2005. <http://eprint.iacr.org/>.

New Security Problem in RFID Systems “Tag Killing”[★]

Dong-Guk Han¹, Tsuyoshi Takagi²,
Ho Won Kim³, and Kyo Il Chung³

¹ Center for Information and Security Technologies(CIST),
Korea University, Seoul, Korea

`christa@korea.ac.kr`

² Future University-Hakodate, Japan

`takagi@fun.ac.jp`

³ Electronics and Telecommunications Research Institute(ETRI), Korea
`{khw, kyoil}@etri.re.kr`

Abstract. Radio frequency identification systems based on low-cost computing devices is the new plaything that every company would like to adopt. The biggest challenge for RFID technology is to provide benefits without threatening the privacy of consumers. Using cryptographic primitives to thwart RFID security problems is an approach which has been explored for several years. In this paper, we introduce a new security problem called as “Tag Killing” which aims to wipe out the functioning of the system, e.g., denial of service attacks. We analyze several well-known RFID protocols which are considered as good solutions with “Tag Killing” adversary model and we show that most of them have weaknesses and are vulnerable to it.

Keywords: Radio frequency identification (RFID), privacy, security, hash chain, challenge-response.

1 Introduction

Often presented as a new technological revolution, Radio Frequency Identification (RFID) makes the identification of objects in open environments possible, with neither physical nor visual contact. RFID systems are made up of transponders inserted into the objects, of readers which communicate with the transponders using radio frequencies and usually of a database which contains information on the tagged objects.

However, these tags also bring with them security and privacy issues. As RFID tag has very low resources: low computing power and small memory size, it is very hard to apply existing security technologies, e.g. asymmetric encryption, that assumes very high computing power and large memory to RFID tag.

[★] This work was done while the first author visits in Future University-Hakodate as Post.Doc.

Up to date, many kinds of protocols have been proposed to resolve RFID privacy problems. In the case of basic tags - lack the resources to perform true cryptographic operations and sometimes rely on hardware techniques - ‘Kill command’ [19], ‘Silent Tree Walking’ [21], ‘Minimalist Cryptography’ [8], ‘Blocker Tag’ [10], ‘Re-encryption’ variants [9, 5, 18], and ‘RFID Guardian’ [15] have been proposed. In the case of smart tags - have richer security capabilities, those capable of computing cryptographic one-way functions including symmetric key encryption, - ‘Ohkubo Type’ [14, 2, 22] and ‘Challenge-Response Type’ [21, 7, 13, 17, 11] are representative.

In this paper, we focus on the security analysis on the previous smart tags. Security problems in RFID systems can be put into two categories. The first concerns those attacks which aim to wipe out the functioning of the system, e.g., denial of service attacks. The second category is related to privacy: the problem is information leakage and also traceability. Even though the first problem is also serious in RFID applications, many researchers have mainly looked into the second problems in order to design protocols which allow authorized persons to identify the tags without an adversary being able to trace them.

We introduce a new security problem called as “Tag Killing” which aims to wipe out the functioning of the system, e.g., denial of service attacks not only by exhausting the resource of a tag, but also by burning out monotonic counters on a tag. It is not the type of privacy related attacks described in [14, 19, 3, 1], but the attack makes a valid tag useless. We analyze several well-known RFID protocols focused on smart tags with “Tag Killing” adversary model and we show that most of them considered as good solutions for privacy [14, 2, 22, 13, 17, 11] have weaknesses and are vulnerable to it. Especially, Kang-Nyang’s protocol [11] was designed to be secure against denial of service attacks, but the security of it is very controversial from our adversary model’s point of view.

In this paper, section 2 gives a brief introduction to RFID systems. Section 3, 4 state assumptions about the security properties of RFID and several security notions, and offers several previous solutions for privacy problems, respectively. In section 5 we propose an new adversary model “Tag Killing” adapted to RFID protocols and we show that most of them have weaknesses and are vulnerable to it. We also sum up the results.

2 RFID Systems

RFID systems are composed of three key elements - Tag, Reader, and Back-end database [21].

2.1 Tags

Every object to be identified in an RFID system is physically labeled with a tag. Tags are typically composed of a microchip for storage and computation, and a coupling element, such as an antenna coil for communication. Tags may also contain a contact pad, as found in smart cards. Tag memory may be read-only, write-once read-many or fully rewritable.

2.2 Readers

Tag readers interrogate tags for their data through an RF interface. To provide additional functionality, readers may contain internal storage, processing power or connections to back-end databases. Computations, such as cryptographic calculations, may be carried out by the reader on behalf of a tag. The channel from reader-to-tag may be referred to as the forward channel. Similarly, the tag-to-reader channel may be referred to as the backward channel.

2.3 Database

Readers may use tag contents as a look-up key into a back-end database. The back-end database may associate product information, tracking logs or key management information with a particular tag. Independent databases may be built by anyone with access to tag contents. This allows unrelated users along the supply chain to build their own applications. It is assumed that a secure connection exists between a back-end database and the tag reader.

3 Security and Privacy Problems

Privacy is one of the most serious problems related to RFID. Designing and analyzing RFID protocols are still a real challenge because no universal model has been defined: up until now designs and attacks have been made in a pedestrian way. In this section, we look around several security notions under the following assumptions.

3.1 RFID Assumptions

1. We assume tag memory is insecure and susceptible to physical attacks [20] revealing their entire contents. The key point is that tags cannot be trusted to store long-term secrets, such as shared keys, when left in isolation.
2. Tag readers are assumed to have a secure connection to a back-end database. Although readers may only read tags from within the short (e.g. 3 meter) tag operating range, the reader-to-tag, or forward channel is assumed to be broadcast with a signal strong enough to monitor from long-range, perhaps 100 meters. The tag-to-reader, or backward channel is relatively much weaker, and may only be monitored by eavesdroppers within the tag's shorter operating range. In this paper, we assume that eavesdroppers may monitor both forward/backward channel without detection.
3. If a tag receives a new request message, while it is already in the authentication session, then the tag should choose its action: to ignore the request or to start immediately a new authentication session for the new query. In this paper, we assume that the tag received a new query starts new authentication session because the former gives a chance for an attacker to lock a tag preemptively.

3.2 Several Security Notions

Indistinguishability (IND). Given a set of readings between tags and readers, an adversary must not be able to find any relation between any readings of a same tag or set of tags. Since tags are not tamper-resistant, an adversary may even obtain the data stored in the memory of the tags additionally to the readings from the readers/tags. It is related with the following privacy problems:

- ID leakage,
- Existing of ID,
- ID tracking, and so on.

Forward Security (FS). Given a set of readings between tags and readers and given the fact that all information stored in the involved tags has been revealed at time t , the adversary must not be able to find any relation between any readings of a same tag or set of tags that occurred at a time $t' \leq t$.

Replay Attack (RA). The eavesdroppers eavesdrop on communication between readers and tags. By eavesdropping, the eavesdropper can take secret information and perform replay attack. The main goal of it is to disguise as the right tags.

4 Several Solutions for RFID Systems

We classify RFID tags according to their computational resources - **basic tags** and **smart tags**. Our categorization is a rough one, of course, as it neglects many other tag features and resources, like memory, communication speed, random-number generation, power, and so forth. It serves our purposes, however, in demarcating available security tools. In this paper, we focus on the problems of privacy and authentication protocols in smart tags.

4.1 Basic Tags

Basic RFID tags lack the resources to perform true cryptographic operations. Basic tags are categorized into two groups, *Pure Tags* and *Tags with Physical Technology*, according to the existence of an additional hardware device to protect consumers from unwanted scanning of RFID.

- **Pure Tags:** ‘Kill Command’ approach [19], ‘Silent Tree Walking’ [21], ‘Minimalist Cryptography’ [8], ‘Re-encryption’ variants [9, 5, 18]¹.
- **Tags with Physical Technology:** ‘Faraday Cage’ [12], ‘Blocker Tag’ [10], ‘Watchdog Tag’ [4], ‘RFID Guardian’ [15].

¹ From several perspectives, like the need for re-encrypting readers, these systems are very cumbersome. But it helpfully introduces the principle that cryptography can enhance RFID tag privacy even when tags themselves cannot perform cryptographic operations.

4.2 Smart Tags

Let us now turn our attention to the class of RFID tags with richer security capabilities, those capable of computing cryptographic one-way functions including symmetric-key encryption, hash function, and so on. We divide smart tags into two categories, *Ohkubo Type* [14, 2, 22] and *Challenge-Response Type* [21, 7, 13, 17, 11], according to the usage of one-way functions.

Ohkubo Type: The Ohkubo type schemes used the hash chain technique to renew the secret information contained in the tag. We introduce the original Ohkubo scheme proposed by Ohkubo-Suzuki-Kinoshita [14].

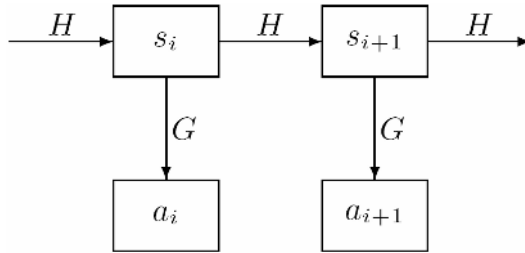


Fig. 1. Illustration of how Ohkubo’s hash chain work

- Initially tag has initial information s_1 . They simply assume that a tag never emits more than n values over its lifetime. The back-end database maintains a list of pairs (ID, s_1) , where s_1 is the initial secret information and is different for each tag. Let H and G be hash functions. In the i -th transaction with the reader,
 - **Reader** Send a query to a tag.
 - **Tag**
 1. sends answer $a_i = G(s_i)$ to the reader,
 2. renews secret $s_{i+1} = H(s_i)$ as determined from previous secret s_i .
 - **Reader** The reader sends a_i to the back-end database.
 - **Database** The back-end database that received tag output a_i from the reader has to identify the corresponding tag. In order to do this, it constructs the hash chains $a'_i = G(H^i(s_1))$ from each initial value s_1 until it finds the expected a_i or until it reaches a given maximum limit n on the chain length.

Property 1. In [14, 2, 22], the lifetime of the tag is a priori limited to n identifications. Note that the size of n is related with the efficiency of performance in the back-end database.

Remarks: The original Ohkubo scheme is proven to be secure against **IND** and **FS** [14]. But it is vulnerable to **RA** [17]. Thus Avoine et al. [2] proposed modified Ohkubo scheme to avoid **RA** using a fresh challenge sent by the reader.

The hash-chain protocols like Ohkubo type have a heavy burden on back-end database to authenticate tags. In order to improve the performance of the back-end servers three approached have been proposed;

- In [14], a more efficient scheme was constructed by making the reader send count number i with a_i and making the back-end server memory store the latest results of hash calculation s_i . By doing so, the performance cost of the back-end server can be cut and facilitate the calculation of a_i for all candidates in the database. Unfortunately, it is not secure against tracking because tag sends the counter number i .
- In [2], Avoine used time-memory trade-off technique [6]. Unfortunately, the time-memory trade-off cannot be applied directly to the modified Ohkubo scheme [2] due to the randomization of the tag’s answer.
- In [22], Yeo-Kim modified the original Ohkubo scheme using the group index in the tag and back-end server. Its security satisfies **IND**, but there is controversial point in **FS** depending on the size of group. Note that **RA** is not guaranteed. The required performance is $O(n\sqrt{m})$ where m is the number of tags and n is the maximum length of the hash chain for each tag.

Challenge-Response Type: Weis et al. and Henrici et al. proposed basic challenge response protocols called “hash-lock protocol” [21] and “Hash-Based ID variation” [7]. However these protocols are vulnerable to **IND**, **FS**, or **RA**. Note that detail security analysis on them is contained in [17].

Molar-Wagner [13] proposed a challenge-response protocol which provides mutual authentication of the reader and the tag. But, Avoine showed that it is not secure against **FS** [2]. Note that **IND** and **RA** are guaranteed.

Recently, Rhee et al. [17] proposed new challenge-response protocol which also provides mutual authentication. The proposed protocol is as follows;

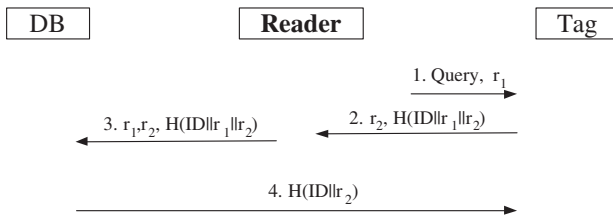


Fig. 2. Protocol of Rhee et al

- **Reader** Send a query and a nonce r_1 to a tag.
- **Tag**
 - The tag picks a random r_2 .
 - Answer $H(ID||r_1||r_2)$ to the reader.
- **Reader** Send $H(ID||r_1||r_2)$ and r_1, r_2 to the database.

- **Database** The database retrieves the identity of the tag by finding $\{ID\}$ in its database. (Authentication of the tag) If the authentication is successful, it sends $H(ID\|r_2)$ to the tag, through the reader.
- **Tag** The tag can thus verify the identity of the reader by checking the received $H(ID\|r_2)$. (Authentication of the reader)

It shall prevent an attacker from **IND** and **RA**, but **FS** may be vulnerable because the tag’s ID is fixed value.

In [11], Kang-Nyang introduced a strong authentication protocol, especially it was designed as a countermeasure against denial of service attack. The basic scheme of it is similar to that of Rhee’s one, but the difference is that the tag changes its ID after authenticating the reader. Thus it seems to be secure against all **IND**, **FS**, and **RA**.

Property 2. In [13, 17, 11], there are common parts to authenticate both a reader and a tag.

1. The reader sends a nonce r_1 with a query to the tag. (The nonce is used for a countermeasure against **RA**.)
2. The tag answers a hashed value with a nonce r_2 generated by itself. The nonce r_2 is stored in the tag.
3. After the checking the validity of the tag’s answer, the reader sends a hashed value including with r_2 , i.e. $H(r_2\|...)$ to the tag. The tag checks the validity of the received hash value which implies the authentication of the reader.

5 New Adversary Model - “Tag Killing”

In this section, we introduce a new adversary model called “Tag Killing” (TK) attack. The goal of it is to stop the service of the valid Tag, i.e. denial of service. It is not the type of privacy related attack, but the attack makes a valid Tag useless. Under Assumption 3 described in session 3.1 if a tag receives a new request message, while it is already in the authentication session, then the tag should open new authentication session.

5.1 How to Kill the Functioning of Tags

The idea is very simple - we assume that there is a valid tag and a invalid reader, i.e. an adversary. The task of the adversary is just to send a lot of queries to the tag. Then the tag should start new authentication session for the newly arrived queries according to the assumption 3. What will happen under this setting?

5.2 A Tag Uses Ohkubo Type Protocols

If the target tag used Ohkubo type protocol then it cannot any more response to any reader after n times answering to the adversary from Property 1. It is

a denial of service attack based on burning out monotonic counters on the tag. Here n is a maximum length of chain. As the size of n is closely related with the performance of the back-end server, simply enlarging the size of n is not so good idea. Thus all Ohkubo type schemes [14, 2, 22] suffer from **TK** because of the limitation of hash chain.

5.3 A Tag Uses Challenge-Response Type Protocols

In the protocols proposed in [13, 17, 11] Property 2 showed that there are common parts authenticating the reader. Among them, we focus on the step 2 in Property 2, i.e. the tag should store the nonce r_2 generated by itself. Due to the limitation of resources in a tag, the allocated memory is very small. Thus, for example, if an adversary sends 2^{20} times queries to the tag then the tag should open new session 2^{20} times which implies that the tag should store 2^{20} random numbers. Assume the size of the random number is 80-bit. Then the required memory size is about 10MB² after receiving 2^{20} queries. The goal of our attack is to make the tag's memory to be exhausted, so let the tag be not able to answer to any query.

Remark 1. Kang-Nyang's protocol [11] was designed to be secure against denial of service attacks, but the security of it is very controversial from our adversary model's point of view because Kang-Nyang's protocol has also the same property 2.

5.4 Toy Example – Shopping

Assumption: In retail shops, consumers could check out by rolling shopping carts past point-of-sale terminals. These terminals would automatically tally the items, compute the total cost. Tags in items are equipped with Ohkubo type or Challenge-Response type scheme.

An adversary selects one item, e.g. an expensive wristwatch, and sends a lot of queries to it until it does not answer. Then he/she puts it into the items he shopped and rolls shopping carts past a point-of-sale terminal. The terminal will automatically tally the items except the wristwatch, computes the total cost which does not contain the cost of the wristwatch. Thus he/she easily steal what he/she want to have without paying the cost if a tag attached to the target item utilizes Ohkubo type or Challenge-Response type protocol as a solution of privacy.

5.5 Summary

We have introduced an new adversary model “Tag Killing” adapted to RFID protocols. We have used this model to analyze the security of many protocols. We sum up the results obtained in Table 1.

² It seems beyond the possible scope of low-cost tags.

Table 1. Analysis of existing RFID protocols: ○- Secure or Satisfactory, ×- Insecure or Unsatisfactory, △- Controversial

Protocol		IND	FS	RA	TK
Ohkubo type	Original Ohkubo scheme [14]	○	○	×	×
	Modified Ohkubo scheme [2]	○	○	○	×
	Ohkubo scheme with grouping [22]	○	△	×	×
Challenge-Response type	Molar-Wagner’s Scheme [13]	○	×	○	×
	Rhee et al’s Schemes [17]	○	×	○	×
	Kang-Nyang’s Scheme [11]	○	○	○	×

Finally, the work presented in this paper is the first step towards discussion of wiping out the functioning of the system, e.g., denial of service attacks against both Ohkubo type and Challenge-Response type.

Up to date, many kinds of papers have discussed the type of privacy related attacks [14, 19, 3, 1], however, our goal of this paper was to propose a new adversary model related with the service of the systems. As we know the “Kill Tag” approach [19, 16] has been considered the most straightforward approach for the protection of consumer privacy, which just kills RFID tags before they are placed in the hands of consumers. There are many environments, however, in which simple measure like “kill” commands are unworkable or undesirable, e.g. theft-protection of belongings, effortless physical access contro, and wireless cash cards. This example shows that applications of good service are also important part considered in RFID besides privacy. As we showed that most of the previous protocols have weaknesses and are vulnerable to our attack, we should try to find a novel solution which is a fully privacy-preserving and does not compromise numerous applications of service.

Acknowledgements

Dong-Guk Han was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

References

1. G. Avoine, *Adversarial model for radio frequency identification*, Cryptology ePrint Archive, Report 2005/049. Referenced 2005 at <http://eprint.iacr.org>.
2. G. Avoine, E. Dysli, and P. Oechslin, *Reducing time complexity in RFID systems*, Selected Areas in Cryptography (SAC 2005), LNCS, Springer-Verlag, 2005. To appear.
3. G. Avoine and P. Oechslin, *RFID traceability: A multilayer problem*, Financial Cryptography (FC 2005), LNCS 3570, pp. 125-140. Springer-Verlag, 2005.
4. C. Floerkemeier, R. Schneider, and M. Langheinrich, *Scanning with a purpose - supporting the fair information principles in RFID protocols*, 2004. Referenced 2005 at citeseer.ist.psu.edu/floerkemeier04scanning.html.

5. P. Golle, M. Jakobsson, A. Juels, and P. Syverson, *Universal reencryption for mixnets*, RSA Conference - CryptographersTrack (CT-RSA 2004), LNCS 2964, pp. 163-178, 2004.
6. M. Hellman *A cryptanalytic time-memory tradeoff*, IEEE Transactions on Information Theory, IT-26:401-406, 1980.
7. D. Henrici and P. Müller, *Hash-based enhancement of location privacy for radio-frequency identification devices using varying identifiers*, Workshop on Pervasive Computing and Communications Security (PerSec 2004), pp. 149-153. IEEE, IEEE Computer Society, 2004.
8. A. Juels, *Minimalist cryptography for low-cost RFID tags*, The Fourth International Conference on Security in Communication Networks (SCN 2004), LNCS 3352, pp. 149-164. Springer-Verlag, 2004.
9. A. Juels and R. Pappu, *Squealing Euros: Privacy protection in RFID-enabled banknotes*, Financial Cryptography (FC 2003), LNCS 2742, pp. 103-121, Springer-Verlag, 2003.
10. A. Juels, R.L. Rivest, and M. Szydło, *The blocker tag: Selective blocking of RFID tags for consumer privacy*, 8th ACM Conference on Computer and Communications Security, pp. 103-111, ACM Press, 2003.
11. J. Kang, and D. Nyang, *RFID Authentication Protocol with Strong Resistance against Traceability and Denial of Service Attacks*, 2nd European Workshop on Security in Ad-Hoc and Sensor Networks, (ESAS 2005), to be appeared.
12. mCloak: Personal / corporate management of wireless devices and technology, 2003. Product description at www.mobilecloak.com.
13. D. Molnar, and D. Wagner, *Privacy and security in library RFID : Issues, practices, and architectures*, ACM Conference on Communications and Computer Security, pp. 210-219, ACM Press, 2004.
14. M. Ohkubo, K. Suzuki, and S. Kinoshita, *Cryptographic approach to "privacy-friendly" tags*, In RFID Privacy Workshop, MIT, USA, 2003.
15. M. Rieback, B. Crispo, and A. Tanenbaum, *RFID Guardian: A battery-powered mobile device for RFID privacy management*, Australasian Conference on Information Security and Privacy (ACISP 2005), LNCS 3574, pp. 184-194, Springer-Verlag, 2005.
16. RFID-Zapper, <https://events.ccc.de/congress/2005/wiki/RFID-Zapper> (EN) .
17. K. Rhee, J. Kwak, S. Kim, and D. Won, *Challenge-Response Based RFID Authentication Protocol for Distributed Database Environment*, International Conference on Security in Pervasive Computing (SPC 2005), LNCS 3450, pp. 70-84. Springer-Verlag, 2005.
18. J. Saito, J.-C. Ryou, and K. Sakurai, *Enhancing privacy of universal re-encryption scheme for RFID tags*, Embedded and Ubiquitous Computing (EUC 2004), LNCS 3207, pp. 879-890, Springer-Verlag, 2004
19. S. E. Sarma, S. A. Weis, and D.W. Engels, *Radio-frequency identification systems*, CHES 2002, LNCS 2523, pp. 454-469, Springer-Verlag, 2002.
20. S.H. Weigart, *Physical Security Devices for Computer Subsystems: A Survey of Attacks and Defences*, In Workshop on Cryptographic Hardware and Embedded Systems, LNCS 1965, pp. 302-317, Springer-Verlag, 2000.
21. S. Weis, S. Sarma, R. Rivest, and D. Engels, *Security and privacy aspects of low-cost radio frequency identification systems*, International Conference on Security in Pervasive Computing (SPC 2003), LNCS 2802, pp. 454-469. Springer-Verlag, 2003.
22. S.S. Yeo and S.K. Kim, *Scalable and Flexible Privacy Protection Scheme for RFID Systems*, 2nd European Workshop on Security in Ad-Hoc and Sensor Networks, (ESAS 2005), to be appeared.

A Model for Security Vulnerability Pattern

Hyungwoo Kang¹, Kibom Kim¹, Soonjwa Hong¹, and Dong Hoon Lee²

¹ National Security Research Institute,
161Gajeong-dong, Yuseong-gu, Daejeon, 305-350, Korea
{kanghw, kibom, hongsj}@etri.re.kr
² Center for Information Security Technologies(CIST),
Korea University, Seoul, 136-704, Korea
Donghlee@korea.ac.kr

Abstract. Static analysis technology is used to find programming errors before run time. Unlike dynamic analysis technique which looks at the application state while it is being executed, static analysis technique does not require the application to be executed. In this paper, we classify security vulnerability patterns in source code and design a model to express various security vulnerability patterns by making use of pushdown automata. On the basis of the model, it is possible to find a security vulnerability by making use of Abstract Syntax Tree (AST) based pattern matching technique in parsing level.

Keywords: Static analysis, Software security, Buffer overflow, Abstract Syntax Tree (AST), Pushdown Automata (PDA).

1 Introduction

Static analysis is extremely beneficial to small and large businesses alike, although the way they are deployed and used may be different. They help keep development costs down by finding bugs as early as possible in the product cycle, and ensure the final application has far fewer exploitable security flaws. Static analysis tools are very good at code level discovery of bugs and can help enforce coding standards and keep code complexity down. Metrics can be generated to analyze the complexity of the code to discover ways to make the code more readable and less complex.

Over the half of all security vulnerabilities, the buffer overflow vulnerability is a single most important security problem. The classic buffer overflow is a result of misuse of string manipulation functions in the standard C library. An example of buffer overflow resulting from misusing *strcpy()* is shown in Figure 1.

```
1 char dst[256];  
2 char *s = read_string();  
3 strcpy(dst, s);
```

Fig. 1. Classic buffer overflow

The string *s* is read from the user on line 2 and can be of arbitrarily long. The *strcpy()* function copies it into the *dst* buffer. If the length of the user string is greater than 256, the *strcpy()* function will write data past the end of the *dst* array. If the array is located on the stack, a buffer overflow can be used to overwrite the return address of a function and execute codes specified by the attacker [1]. However, the mechanics of exploiting software vulnerabilities are outside the scope of this work.

A number of static analysis techniques have been used to detect specific security vulnerabilities in software. Most of them are not suitable for large scale software. In this paper, we propose a new mechanism being able to detect security vulnerability patterns in large scale source codes by making use of compiler technologies.

This paper is organized as follows. Chapter 2 reviews related researches. In chapter 3, a new mechanism for checking security vulnerability patterns is introduced. The implementation and experiment on proposed mechanism are showed in chapter 4. Finally, in chapter 5, we draw conclusions.

2 Related Works

Previous static analysis techniques can be classified into the following two types: lexical analysis based approach and semantic based approach.

2.1 Lexical Analysis Based Approach

Lexical analysis technique is used to turn the source codes into a stream of tokens, discarding white space. The tokens are matched against known vulnerability patterns in the database. Well-known tools based on lexical analysis are Flawfinder [2], RATS [3], and ITS4 [4]. For example, these tools find the security vulnerability of *strcpy()* by scanning security-critical source codes in Figure 1. While these tools are certainly a step up from UNIX utility *grep*, they produce a hefty number of false positives because they make no effort to account for the target code's semantics. A stream of tokens is better than a stream of characters, but it's still a long way from understanding how a program will behave when it executes.

To overcome the weakness of lexical analysis approach, a static analysis tool must leverage more compiler technology. By building an abstract syntax tree (AST) from source code, such a tool can take into account the basic semantics of the program being evaluated. Lexical analysis based tools can be confused by a variable with the same name as a vulnerable function name, but AST analysis will accurately distinguish the different kinds of identifiers. On the AST level, complicated expressions are analyzed, which can reveal vulnerabilities hidden from lexical analysis based tools.

2.2 Semantic Based Approach

BOON [5] applies integer range analysis to detect buffer overflows. The range analysis is flow-insensitive and generates a very large number of false alarms. CQUAL [6] is a type-based analysis tool that provides a mechanism for specifying and checking properties of C programs. It has been used to detect format string vulnerability. SPLINT [7] is a tool for checking C program for security vulnerabilities and programming mistakes. It uses a lightweight data-flow analysis to verify assertions about

the program. Most of the checks performed by SPLINT require the programmer to add source codes annotations which guide the analysis.

An abstract interpretation approach [8, 9] is used for verifying the Airbus software. The basic idea of abstract interpretation is to infer information on programs by interpreting them using abstract values rather than concrete ones, thus, obtaining safe approximations of programs behavior. Although this technique is a mature and sound mathematical approach, the static analysis tool which uses abstract interpretation can't scale to larger code segments. According to the testing paper[10], target program has to be manually broken up into 20-40k lines-of-code blocks to use the tool. So, we need a static analysis approach to handle large and complex programs.

Microsoft set up the SLAM [11, 12, 13] project that uses software model checking to verify temporal safety properties in programs. It validates a program against a well designed interface using an iterative process. However, SLAM does not yet scale to very large programs because of considering data flow analysis. MOPS [14, 15] is a tool for verifying the conformance of C programs to rule that can be expressed as temporal safety properties. It represents these properties as finite state automata and uses a model checking approach to find if any insecure state is reachable in the program. MOPS, however, isn't applicable to various security rules because of considering order constraint only. A lot of security vulnerability can not be expressed by temporal safety properties. Therefore, a new static analysis technique that is possible to express various security properties is needed.

3 Mechanism for Checking Security Vulnerability Pattern

In this chapter, we propose a mechanism for checking security vulnerability patterns in source codes. The mechanism uses an AST based pattern matching and PDA in order to find vulnerability pattern in target source codes.

3.1 Problem

There are lots of security vulnerability patterns in source codes causing system crash or illegal system privilege acquisition. We classify the vulnerability patterns in source level into following 3 types.

Vulnerability pattern type 1

Type 1 is the simplest pattern having only one function, such as *strcpy()* or *strcat()*. Figure 1 is a typical example for vulnerability pattern type 1. This vulnerability type is easily detected by lexical analysis tools which use pattern matching for single token in code level. In this case, these tools can detect the security vulnerability using *strcpy()* as a token.

Vulnerability pattern type 2

Type 2 is a pattern providing order constraint which is expressed by more than two tokens, such as function names. For example, Figure 2(a) shows an insecure program having pattern type 2. A setuid-root program should drop root privilege before execut-

ing an untrusted program. Otherwise, the untrusted program may execute with root privilege and therefore compromise the system.

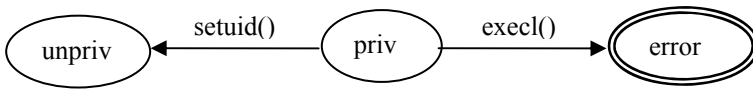
The pattern type 2 can be detected by MOPS or SLAM by making use of model checking technique using finite state automata (FSA) to checking temporal safety property. Figure 2(b) shows simplified FSA describing `setuid-root` vulnerability.

```

1 // The program has root privilege
2 if ((passwd = getpwuid(getuid())) != NULL) {
3     fprintf(log, "drop priv for %s", passwd->pw_name);
4     setuid(getuid()); // drop privilege
5 }
6 execl("/bin/sh", "/bin/sh", NULL); // risky syscall

```

(a) `setuid-root` program



(b) A FSA describing `setuid-root` vulnerability

Fig. 2. Security vulnerability pattern type 2

Vulnerability pattern type 3

Type 3 is the most complex pattern including function names, operators, and variables. Figure 3 shows an example of MS windows Bof vulnerability published in 2003[16, 17]. The program is a typical program having vulnerability pattern type 3.

```

1 // buffer copy while the condition is TRUE
2 while ( *id2 != '\')
3     *id1++ = *id2++; // buffer copy

```

Fig. 3. Example of security vulnerability pattern type 3

The program in Figure 3 copies from a buffer id_2 to another buffer id_1 while the condition on line 2 is true. There is no consideration about the size of target buffer while buffer copy. There is no problem in the program at ordinary times. However, buffer overflow happens when an attacker fills the buffer id_2 with no `\` character up to size of target buffer id_1 . The Bof vulnerability in Figure 3 can't be detected by previous model checking tool, such as MOPS and SLAM, because these tools consider order constraints only. But the source codes may have the pattern type 3 vulnerability in real environment. Therefore, a new mechanism being able to detect vulnerability pattern type 3 is needed.

We call an expressional safety property to detect vulnerability pattern type 3. An expressional safety property dictates the set of various security-relevant operations including function names, operators, and variables.

3.2 Model of Security Vulnerability Pattern

We introduce a new model for expressional safety property in order to detect pattern type 3 including various tokens such as functions, operators, and variables. The model is based on parse tree (AST) which is a structural representation of sentences (expressions) in program source code.

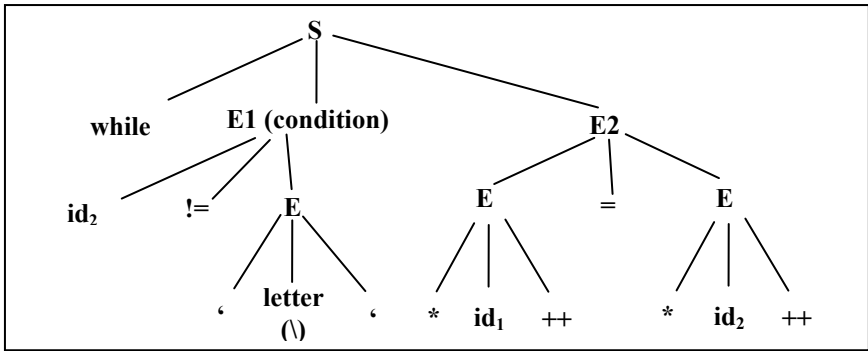


Fig. 4. Parse tree describing the program in Figure 3

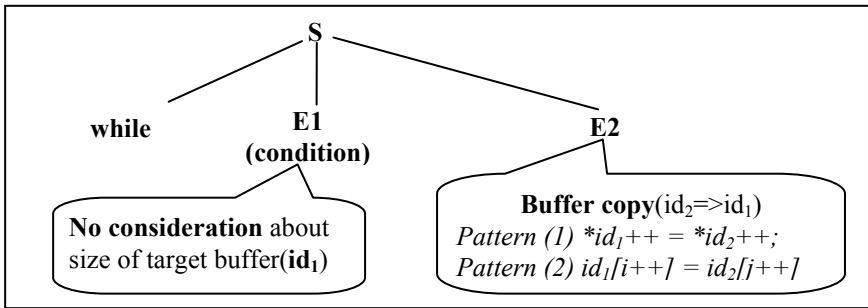


Fig. 5. Simplified parse tree for original tree in Figure 4

Figure 4 shows a parse tree which is output of parsing for the program in Figure 3 having Bof vulnerability. The S stands for a sentence and the E stands for an expression.

We can simplify the parse tree to an abstracted model, such as Figure 5. The E1 means that there is no consideration about the size of target buffer id_1 . Any security vulnerability doesn't exist when there is a consideration about the size of target buffer id_1 in E1. Therefore, we simply check whether there is a target buffer id_1 in E1 or not.

E2 means that there is a copy statement from source buffer id_2 into the target buffer id_1 . We can check whether there is the pattern (1) or (2) in E2 or not. The Parse tree can be recognized by context-free Grammar (CFG) which is type 2 grammar in Chomsky Hierarchy. A CFG takes a PDA to recognize context-free language. That means we need to construct PDA in order to recognize the security vulnerability pattern type 3. Figure 6 shows a PDA to recognize the parse tree in Figure 5.

$$\begin{aligned}
 \text{PDA } M = (Q, \Sigma, \Gamma, \delta, q_0, Z, F), \text{ where } & Q = \{q_0, q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8\}, \\
 & \Sigma = \{\text{while}, id_1, id_2, *, =, ++, [,]\}, \\
 & \Gamma = \{Z, id_1, id_2\}, F = \{q_8\} \\
 \\
 \delta(q_0, \text{while}, Z) = \{(q_1, Z)\}, & \delta(q_1, id_2, Z) = \{(q_2, id_2Z)\} \\
 \delta(q_2, *, id_2) = \{(q_3, id_2)\}, & \delta(q_3, id_1, id_1) = \{(q_0, \varepsilon)\} \\
 \delta(q_3, id_1, id_2) = \{(q_4, \varepsilon)\}, & \delta(q_4, ++, Z) = \{(q_5, Z)\} \\
 \delta(q_5, *, Z) = \{(q_6, Z)\}, & \delta(q_6, id_2, Z) = \{(q_7, Z)\} \\
 \delta(q_7, ++, Z) = \{(q_8, \varepsilon)\} &
 \end{aligned}$$

Fig. 6. PDA M being able to recognize simplified parse tree in Figure 5

The reason why PDA is needed to recognize security vulnerability pattern type 3 is that FSA does not have stack storage. Stack is used to store the variable id_2 in E1 and to extract the variable id_2 in E2 at Figure 4. To check the vulnerability in the program, we need to check whether there exists the id_1 in E1 when E2 is processed. Therefore, PDA having stack storage is needed. The program in Figure 3 has vulnerability because there is no variable id_1 in condition of while loop at Figure 4. That means the PDA M using stack storage recognizes the program in Figure 3. After all, a target program has a vulnerability violating expressional safety property when the program reaches a final state (that is an error state) of PDA.

We model the expressional safety property as a PDA M. Then, we use AST based pattern matching to determine whether a state violating the expressional safety property is reachable in the target program. We can check for various security properties by making use of AST based pattern matching technique in parsing level. This approach determines in compile time whether there are any security vulnerability patterns in target program that may violate a security property. So, proposed mechanism is suitable for checking security property for large scale software.

Our model can express not only vulnerability pattern type 3 but also vulnerability pattern type 1 and 2 which can be expressed by FSA. According to Chomsky Hierarchy, a set of context-free language recognized by PDA includes a set of regular language recognized by FSA.

3.3 Formal Expression

We present a formal mechanism for checking expressional safety property. Let Σ be the set of security-relevant operations. Let $B \subseteq \Sigma^*$ be a set of sequences of security operations that violate the security property (B stands for bad). An expression $e \in \Sigma^*$

will represent a sequence of operations expressed in target program. Let $E \subseteq \Sigma^*$ denote the set of all feasible expressions, extracted from all statements of the program (E stands for expression). The problem is to decide if $E \cap B$ is empty. If so, then the security property is satisfied. If not, then some expressions in the program may violate the security property.

In the above model, B and E are arbitrary languages. First, we showed that B , the set of sequence of security operations that violate the expressional safety property, is a context-free language. The reason why B is not a regular language is that pushdown automaton is needed to recognize various vulnerability patterns including functions, operators, and variables. Especially, temporal storage, such as stack, is used to memorize the variables. We show that most expressional safety properties can be described by context-free languages (see Sections 3.2). Since B is a context-free language, there exists a PDA M that accepts B (M stands for model); in other words, $B = L(M)$.

We need to show if $L(M) \cap E$ is empty. Since E is the set of all feasible expressions in target program, we have to check whether the set E is accepted by a PDA M . If the PDA M accepts the E of feasible expression in target program, there is a security vulnerability in the target program violating the security property defined in advance.

According to automata theory, there are efficient algorithms to determine if a language is accepted by a PDA[18]. Hence we obtain a means to verify whether the security property is satisfied by the program.

We are guaranteed that $E \cap B = E \cap L(M)$. Consequently, if $E \cap L(M)$ is empty, we can conclude that $E \cap B$ is also empty, hence the program definitely satisfies the security property; in contrast, if $L(M) \cap E$ is non-empty, then we can only say that $E \cap B$ is non-empty, hence the program may not satisfy the security property. This means that our analysis is sound and there is no false negative in proposed mechanism. However, in case of making use of inappropriate expressional safety property, a false positive could be happened in proposed mechanism. So, it is very important to make use of well-defined expressional safety property in proposed mechanism.

3.4 Identification of Vulnerability

The Figure 7 shows a concrete process recognizing security vulnerability based on proposed mechanism. The problem is to check whether the target program in Figure 3 violates the security property having a buffer copy without consideration about size of

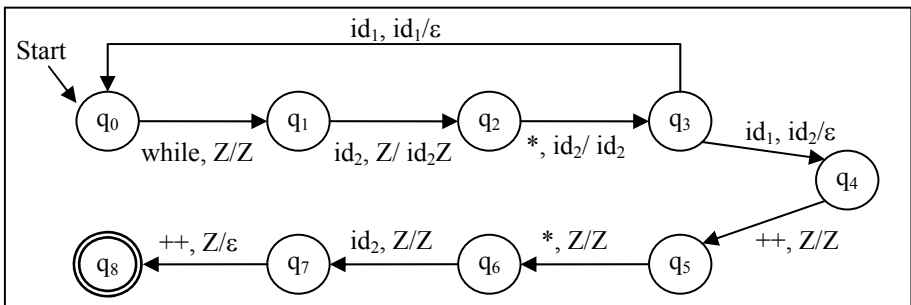


Fig. 7. Process of identification for security property modeled by PDA M in Figure 5

target buffer. In this problem, the set of security operations is $\Sigma = \{\text{while}, \text{id}_1, \text{id}_2, *, =, ++, [,]\}$. The set $B \subseteq \Sigma^*$, the sequences of security-relevant operations that violate the security property, is accepted by the PDA M shown in Figure 6. The set $E \subseteq \Sigma^*$, the feasible expressions of the program in Figure 3, is $E = \{\text{[while, id}_2, *, \text{id}_1, ++, =, *, \text{id}_2, ++], \dots\}$. According to Figure 3, the sequence $[\text{while, id}_2, *, \text{id}_1, ++, =, *, \text{id}_2, ++]$ in E is accepted by PDA M. Therefore, we find that $E \cap L(M) \neq \emptyset$, or in other words, we can recognize the existence of an expression in the target program which violates the security property. This indicates the presence of security vulnerability.

4 Implementation and Experiment

In this chapter, we present the implementation and results of experiment on proposed mechanism.

4.1 Implementation

The implemented tool consists of a parser and a model checker. The parser takes a C source codes as a target program and outputs its AST. The model checker takes the AST and PDA describing an expressional safety property, and decides if any expression in the target program may violate the security property. If so, the tool reports these expressions. The Figure 8 shows a brief architectural overview of implemented tool which checks security vulnerability in source code. Our tool is implemented as a module in CIL[19], a front end for C written in OCaml[20]. CIL parses C codes into an AST format and provides a framework for performing passes over this AST.

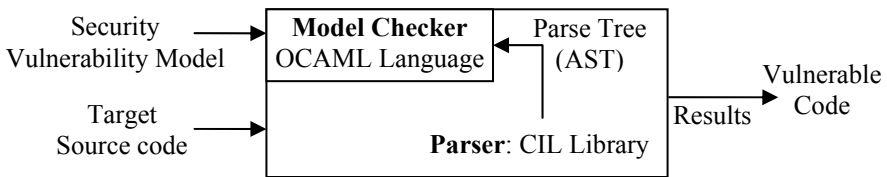


Fig. 8. Architecture of implemented tool for checking expressional safety property

4.2 Result of the Experiment

We make an experiment on 2 source programs in which have security vulnerability shown in Figure 3. Target programs are sample source codes including critical security vulnerability in WebDav[16](Figure 9) and RPCSS Service[17](Figure 10) respectively. The second vulnerability in RPCSS Service is famous as Blaster Worm.

We model the security vulnerability in 2 target programs to PDA M shown in Figure 6. The codes of Bold and italic type in Figure 9 and 10 have pattern of security vulnerability which is modeled to PDA M. The implemented tool succeeds in detecting security vulnerability in target programs.

```

long PathName(long nBufferLength, char *Buffer2){
    while ( *Buffer2 && !IS_HAVE_SEPERATOR(*Buffer2) ) {
        if ( *Buffer2 == ';' ) {
            Buffer2++;
            *Buffer1++ = *Buffer2++;
        }
        else *Buffer1++ = *Buffer2++;
    }
}

```

Fig. 9. First target source codes having security vulnerability in WebDav

```

int GetComputerName( char *InputBuffer, char MachineName[MAX_LENGTH] ){
    char *Buffer2 = InputBuffer + 2;
    while ( *Buffer2 != ';' )
        *Buffer1++ = *Buffer2++;
}

```

Fig. 10. Second target source codes having security vulnerability in RPCSS (Blaster Worm)

5 Conclusion

Static analysis is a proven technology in the implementation of compilers and interpreters. Recently, it has begun to see application of static analysis techniques in novel areas such as software validation and bug checking in software. We conclude by summarizing the main contributions of our work:

- We classify the security vulnerability patterns into 3 types. Type1 is simple pattern having suspicious single token. Type 2 has more than 2 tokens being able to consider temporal safety properties. Type 3 is the most complex pattern being able to consider various expressional safety properties.
- A model expressing various security vulnerability patterns is provided. The model is used to express various expressions related security property when the proposed mechanism checks a target program for security vulnerability. All of the vulnerability pattern types can be expressed by our model. There was no attempt to model these vulnerability types.
- A mechanism based on formal language theory is proposed. The mechanism provides an infrastructure being able to check a target program for security properties which is defined by users in advance.
- We implemented proposed mechanism as a tool for checking expressional safety property. As we mentioned in the section 4.2, the tool showed excellent results in detecting security vulnerability.

The proposed mechanism has several advantages:

- It is sound because it can reliably detect all bugs of the specified properties.
- It can check for various security vulnerability patterns by making use of PDA.
- It is scalable due to using AST based pattern matching technique in parsing level.

References

1. Aleph One: Smashing the stack for fun and profit. Phrack 49-14 (1996)
2. D. A Wheeler: Flawfinder. <http://www.dwheeler.com/flawfinder/>.
3. RATS. <http://www.securesw.com/rats/>.
4. J. Viega, J. T. Bloch, T. Kohno and G. McGraw: ITS4: A static vulnerability scanner for C and C++ code. *ACM Transactions on Information and System Security* 5(2) (2002).
5. D. Wagner, J. S. Foster, E. A. Brewer and A. Aiken: A first step towards automated detection of buffer overrun vulnerabilities. In *Network and distributed system security symposium*, 3–17. San Diego, CA (2000)
6. J. Foster: Type qualifiers: Lightweight specifications to improve soft-ware quality. Ph.D. thesis, University of California, Berkeley (2002)
7. D. Evans: SPLINT. <http://www.splint.org/>.
8. B. Blanchet, P. Cousot, R. Cousot, J. Feret, L. Mauborgne, A. Mine, D. Monniaux and X. Rival: A Static Analyzer for Large Safety-Critical Software (2003)
9. Abstract interpretation. <http://www.polyspace.com/downloads.htm> (2001)
10. M. Zitser, R. Lippmann, T. Leek: Testing Static Analysis Tools using Exploitable Buffer Overflows from Open Source Code, pp.97-106, SIGSOFT'04 (2004)
11. T. Ball, R. Majumdar, T. Millstein, and S. Rajamani: Automatic predicate abstraction of C programs. *PLDI. ACM SIGPLAN Not.* 36(5) (2001), 203–213.
12. T. Ball, A. Podelski, and S. Rajamani: Relative completeness of abstraction refinement for software model checking. *TACAS (2002)*, LNCS 2280, Springer, 158–172.
13. T. Ball and S. Rajamani: The SLAM project: debugging system software via static analysis. *29th ACM POPL (2002)*, LNCS 1254, Springer, 72–83.
14. H. Chen and D. Wagner: MOPS: an infrastructure for examining security properties of software. In *Proceedings of the 9th ACM Conference on Computer and Communications Security (CCS)*, Washington, DC (2002)
15. H. Chen, D. Wagner, and D. Dean: Setuid demystified. In *Proceedings of the Eleventh Usenix Security Symposium*, San Francisco, CA (2002)
16. Microsoft Security Bulletin MS03-007. <http://www.microsoft.com/technet/security/bulletin/MS03-007.msp>. Microsoft (2003)
17. Microsoft Security Bulletin MS03-026. <http://www.microsoft.com/technet/security/bulletin/MS03-026.msp>. Microsoft (2003)
18. J. Hopcroft and J. Ullman: *Introduction to automata theory, languages, and computation*. Addison-Wesley (1979)
19. G. C. Necula, S. McPeak, S. P. Rahul and W. Weimer: CIL:Intermediate Language and Tools for Analysis and Transformation of C Programs. In *Proceedings of CC 2002: 11'th International Conference on Compiler Construction*. Springer-Verlag, Apr. 2002.
20. D. R'emy and J. Vouillon: Objective ML: An effective object-oriented extension of ML. *Theory and Practice of Object Systems*, 4(1):27–52, 1998.

A New Timestamping Scheme Based on Skip Lists*

Kaouthar Blibech and Alban Gabillon

LIUPPA/CSySEC,
Université de Pau – IUT de Mont de Marsan, France
k.blibech@etud.univ-pau.fr
alban.gabillon@univ-pau.fr

Abstract. Time stamping is a cryptographic technique providing us with a proof-of-existence of a message/document at a given time. Several timestamping schemes have already been proposed [1-10]. In this paper, we first define a new timestamping scheme which is based on skip lists [11]. Then, we show that our scheme offers nice properties and optimal performances.

1 Introduction

Timestamping is a technique for providing proof-of-existence of a message/document at a given time. Timestamping is mandatory in many domains like patent submissions, electronic votes or electronic commerce. Timestamping can ensure non-repudiation. Indeed, a digital signature is only legally binding if it was made when the user's certificate was still valid, and a timestamp on a signature can prove this. Parties of a timestamping system are the followings:

Client: Forms the *timestamping request* which is the *digest* of the document to be timestamped. The client computes this digest by using a well known one-way¹ collision-free² hashing function. Submitting the digest of the document instead of the document itself preserves the confidentiality of the document.

TimeStamping Authority (TSA): Receives the timestamping request at time t and issues the *timestamp*. The timestamp is a proof that the digest was received at time t . The TSA produces the timestamp according to a *timestamping scheme*.

Verifier: Verifies the correctness of the timestamp by using the *verification scheme* corresponding to the timestamping scheme which was used to produce the timestamp.

Most of the existing timestamping schemes are *linking* schemes. Linking schemes were introduced by Haber and Stornetta [7]. Such schemes significantly reduce the scope of operations the TSA has to be trusted for. Basically, they work as follows:

* This work was supported by the Conseil Général des Landes and the French ministry for research under *ACI Sécurité Informatique 2003-2006, Projet CHRONOS*.

¹ One-way means that no portion of the original document can be reconstructed from the digest.

² Collision-free means that it is infeasible to find x and x' satisfying $h(x) = h(x')$.

During a time interval which is called a *round*, the TSA,

- receives a set of timestamping requests,
- aggregates the requests in order to produce a *round token*,
- returns the timestamps to the clients. Each timestamp consists of the round token, the digest and the authentication path proving that the round token depends on the digest.

Each round token is one-way dependent on the round tokens issued before. Round tokens are regularly published in a widely distributed media (a newspaper). After the publication it becomes impossible to forge timestamps (either to issue fake ones afterwards, or modify already issued ones), even for the TSA.

In the case of *partially ordered linking schemes* [1][2][3], only timestamps from different rounds are comparable whereas in the case of *totally ordered linking schemes* [5][6][7], the temporal order of any two timestamps can be verified even if these two timestamps belong to the same round. Partially ordered schemes are generally simpler than totally ordered schemes. However, as mentioned by Arne et al. [12], since totally ordered linking schemes allow us to compare two timestamps of the same round, longer rounds can be used. Using longer rounds enables reducing the amount of data to be published and the amount of data to be verified.

The purpose of this paper is to define a new totally ordered scheme which is simpler than the existing ones and which shows optimal performances. Our scheme uses a skip list. A skip list is a data structure which was defined by Pugh [11].

This paper is organized as follows: section 2 reviews related works. Section 3 presents our scheme. Section 4 deals with performance issues. Finally section 5 concludes this paper.

2 Related Works

Our scheme can be compared to the following existing schemes:

- Partially ordered timestamping schemes [1] [2][3]
- Totally ordered timestamping schemes [5][6][7]

Most of the existing partially ordered timestamping schemes are either based on Merkle trees (binary trees) [1][2] or on cryptographic accumulators [3]. With these schemes, only timestamps from different rounds are comparable. Moreover, schemes based on Merkle trees require the number of requests per round to be a power of 2 whereas schemes based on accumulators generally introduce a cryptographic trapdoor due to the use of the RSA modulus.

Existing totally ordered timestamping schemes are the simply linking scheme [7], the binary linking scheme [5] and the threaded authentication tree scheme [6]. The verification procedure for the simply linking scheme is costly ($O(n)$, where n is the number of received requests) and requires that the TSA saves the entire chronological chain of timestamps.

The binary linking scheme uses a simply connected authentication graph. In addition to its complexity, this scheme is less efficient in terms of time complexity than the Merkle tree scheme for both timestamping and verification due to additional concatenation operations.

The threaded tree scheme can be seen as an improvement of the Merkle tree scheme. It is easier to implement than the binary linking scheme and it issues smaller timestamps. However, when compared to other schemes based on Merkle trees, it still has larger time complexity for both timestamping and verification due to the additional concatenation operations.

3 A New Timestamping Scheme

3.1 Skip Lists

W. Pugh introduced skip lists as an alternative data structure to search trees [11]. The main idea is to add pointers to a simple linked list in order to skip a large part of the list when searching for a particular element. While each element in a simple linked list points only to its immediate successor, elements in a skip list can point to several successors.

Skip lists can be seen as a set of linked lists, one list per level (see figure 1). All the elements are stored at the first level 0. A selection of elements of level k is included in the list at level $k+1$. In *probabilistic* skip lists, if element e belongs to level k then it belongs to level $k + 1$ with probability p . In *deterministic* skip lists, if element e belongs to level k and respects some given constraints, then it belongs to level $k+1$. For example, in *perfect* skip lists (see figure 1), which are the most known deterministic skip lists, element e belongs to level i if its index is a multiple of 2^i . Consequently, element at index 5 belongs only to the first level, while element at index 4 belongs to the three first levels. In figure1, B and E nodes are stored at all levels and called *sentinel* elements. The highest B node is called *starting node* S_t . The highest E node is called *ending node* E_t .

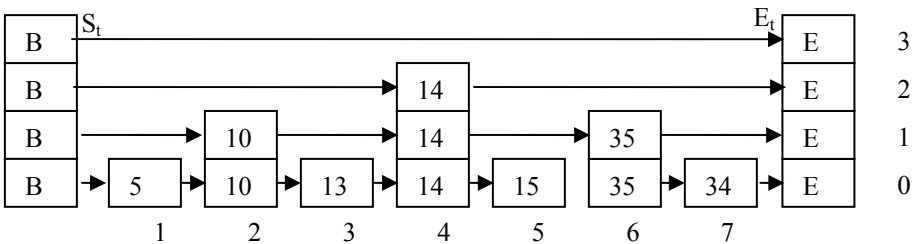


Fig. 1. Nodes contain the elements of the set {5,10,13,14,15,35,34}. Edges are the pointers. Numbers [0..3] are the levels. Numbers [1..7] are the indexes.

3.2 Timestamping Scheme

In [13], we defined an *authenticated dictionary* based on skip lists. An authenticated dictionary is a data structure that supports both update queries and *tamper-evident* membership queries. A tamper-evident membership query is of the form “does element e belong to set S ?”. If e belongs to S then the answer to such a query is a proof that e belongs to S .

The purpose of this paper is to define a new totally linking timestamping scheme based on the dictionary we defined in [13]. Our scheme uses one *append-only* perfect skip list per round. Elements of the skip lists are the timestamping requests. Each new request is appended to the skip list. Since we are dealing with perfect skip lists, each element of the skip list is associated to one or several nodes according to the index of the request. Each node has the following four properties:

- Its *value*, which is a timestamping request (digest)
- Its *level*, ranging from 0 to the highest level of the skip list
- Its *index*, which is its position in the skip list
- Its *label*, which is a hash value one way dependent on the labels of the previous nodes.

Nodes associated to the same element have the same value and index. For example, let us consider nodes a and p in figure 2. They have the same index (20) and value (h_{20}). Level of node a is 2 whereas level of node p is 0. Labels of nodes a and p are not shown but they are different from each other.

The label of the starting node is the round token of the previous round whereas its value is the last request which was received during the previous round. Basically, our scheme works as follows:

- Alice sends a timestamping request which is the digest h of a document.
- The TSA appends h to the skip list.
- The TSA immediately returns to Alice a signed *acknowledgment* containing the index of h in the skip list and the proof that h is inserted after the elements which are already in the skip list. We call this proof the *head proof* (see algorithm 1).
- The TSA computes the label of each node associated to element h (see algorithm 2).
- At the end of the round, the TSA inserts the last request which becomes the ending sentinel element. The label of the ending node is the round token.
- The TSA publishes the round token and sends to Alice (and other clients) some additional information allowing her to prove that her request belongs to the round whose token has been published. We call this information the *tail proof* (see algorithm 3). The final timestamp consists of the digest h , the index of h , the head proof, the tail proof and the round token.
- If a verifier, Bob, needs to check the validity of the timestamp then he has to verify that he can compute the round token from h , the index of h , the head proof and the tail proof. Bob does the verification by processing algorithm 4.

Figure 2 shows the insertion of h_{21} at index 21. h_{16} to h_{21} are requests (digests of documents). Numbers [16..21] are indexes. Labels are not shown. The arrows denote the flow of information for computing the labels (see algorithm 2). The head proof for h_{21} consists of the labels of the dark grey nodes (nodes q , o and a) (see algorithm 1).

Figure 3 shows the insertion of the ending node (last request of the round). The arrows denote the flow of information for computing the labels (see algorithm 2). The label of the ending node is the round token. The tail proof for h_{21} consists of the value h_{22} and the labels of the light grey nodes (nodes r and x) (see algorithm 3). Note that the last request of the round is h_{25} . Since it is the ending element, it belongs to all levels although 25 is not a multiple of 2^5 . Figure 3 shows also the verification process

of node q (that will be used during the verification to compute the label of the ending node i.e. the round token).

Algorithm 2 is used to compute the labels of the nodes associated to the newly inserted element h . Function $value(n)$ returns the value of node n . Function $left(n)$ returns the left node of node n . For example, the left node of node t is node a (see figure 3). Function $down(n)$ returns the bottom node of node n . For example, the bottom node of node d is node c (see figure 3). $hash$ is a one-way collision-free hashing function and \parallel is the concatenation operation. Algorithm 2 applies to each node associated to the newly inserted element starting from the node at level 0 until the plateau node.

Algorithm 2. Hashing Scheme

```

1: If  $down(n) = null$ ,  $\{n$  is at level 0 $\}$ :
2:   If  $left(n)$  is not a plateau node then
3:      $label(n) := value(n)$ .                                { case 1 }
4:   Else
5:      $label(n) := hash(value(n) \parallel label(left(n)))$     { case 2 }
6: Else :
7:   If  $left(n)$  is not a plateau node then
8:      $label(n) := label(down(n))$                             { case 3 }
9:   Else
10:     $label(n) := hash(label(down(n)) \parallel label(left(n)))$  { case 4 }

```

Let us consider node r in figure 3 (index 24, value h_{24} and level 0) and node e (index 23, value h_{23} and level 0). The label of node r is equal to the hash of the value of node r (h_{24}) concatenated to the label of node e (case 2). Now, let us consider node s (index 24, value h_{24} and level 1) and node d (index 22, value h_{22} and level 1). The label of node s is equal to the hash of the label of node r concatenated to the label of node d (case 4). Let us consider also node b (index 21, value h_{21} and level 0) and node p (index 20, value h_{20} and level 0). Node p is not a plateau node, so the label of node b is equal to its value h_{21} (case 1). Finally, let us consider node d (index 22, value h_{22} and level 1) and node c (index 22, value h_{22} and level 0). The label of node d is equal to the label of node c (case 3).

Algorithm 3 is used to compute the tail proof (tp) of elements which were inserted during the round. Function $right(n)$ returns the right node of node n . For example, the right node of node d is node s (see figure 3). Function $top(n)$ returns the node on top of node n . For example, the top node of node s is node t (see figure 3). Computation of the tail proof of element h starts from the plateau node associated to element h (in algorithm 3, n is initialized to the plateau node of element h).

Algorithm 3. Tail proof computation

$\{n$ is initialized to the plateau node of the element $\}$

```

1:  $tp := \{\}$ 
2: While  $right(n) \neq null$  :
3:    $n := right(n)$ 
4:   if  $down(n) = null$  then
5:     append  $value(n)$  to  $TP$ 
6:   Else
7:     append  $label(down(n))$  to  $TP$ 
8:   While  $top(n) \neq null$  :
9:      $n := top(n)$ 

```

Figure 3 shows that the tail proof of element h_{2l} consists of h_{22} (that will be used during the verification to compute the label of node c), the label of node r (that will be used during the verification to compute the label of node s) and the label of node x (that will be used during the verification to compute the label of node y).

3.3 Verification Scheme

We call the *traversal chain* of element h the smallest sequence of labels that have to be computed from h in order to determine the round token (label of the ending node Et). An example of such a chain is given by the labels of the thick nodes in Figure 3. They represent the traversal chain of element h_{2l} . The final timestamp consists of the digest h , the index of h , the head proof of h , the tail proof of h and the round token. It contains all the necessary information to compute the traversal chain of h . The verification process succeeds if the last label of the computed traversal chain is equal to the round token. If not, the verification fails.

Algorithm 4 describes the verification process. Regarding that algorithm, we need to define the following functions:

- $height(index)^3$ that returns the level of the plateau node at position $index$
- $leftIndex(index, level)^4$ that returns the index of the left node of node of index $index$ and level $level$
- $hasPlateauOnLeft(index, level)^5$ that indicates if the node of index $index$ and level $level$ has a plateau node on its left.
- $getNext()$ that extracts the next label from the tail proof,
- $getPrec()$ that extracts the next label from the head proof,
- $getNextIndex(index)^6$ that returns the index of the node whose label is the next label to be extracted by $getNext()$. That index can be computed from $index$.

In algorithm 4, h denotes the request and i_h the index of h (included in the timestamp). $token$ denotes the round token included in the timestamp. Variable $label$ denotes the label of the *current* node in the traversal chain. It is initialized to h .

As we can see, the index of the request in the skip list is a parameter of algorithm 4. If the TSA would lie on the index of the request, then the verification would fail since it would not be possible to compute the labels of the nodes belonging to the traversal chain. Since the head proof is returned as a signed acknowledgement immediately after the request was received, the TSA cannot reorder elements in the skip list even before publishing the round token.

³ Since we are dealing with perfect skip lists, the height h of any element can be computed from its index i : $i = 2^h * k$ where $HCF(2, k) = 1$.

⁴ Since we are dealing with perfect skip lists, the left node of a node of index i and level l has an index $j = i - 2^l$.

⁵ Consider node n of index i and level l . Consider k such that $i - 2^l = k * 2^l$. Since we are dealing with perfect skip lists, if $HCF(2, k) = 1$ then the left node of n is a plateau node.

⁶ Since we are dealing with perfect skip lists, the next index j can be computed from the current index i : $j = i + 2^h$, where h is the height of the element at position i .

Algorithm 4. Verification process

```

1 : {h is the request,  $i_h$  the index of h}
2 : label := h
3 : index :=  $i_h$ 
4 : level := 0
5 : While TP != {}
6 :     For i from level to height(index) :
7 :         If hasPlateauOnLeft(index, i) then
8 :             If leftIndex(index, i) <  $i_h$  then
9 :                 label := hash(label||getPrec())
10:            If leftIndex(index, i) ≥  $i_h$  then
11:                label := hash(getNext())||label
12:            level := i.
13:            index := getNextIndex(index).
14: While HP != {} :
15:     label := hash(label||getPrec()).
16: If label = token then return TRUE
17: Else return FALSE

```

Figure 3 shows the verification process for h_{21} ($i_{h_{21}} = 21$). Labels of thick nodes are computed during the verification process. Variable *label* is initialized to h_{21} . Initial node of level 0 and index 21 is node *b*. Index of left node of node *b* is $21-2^0 (=20)$. Left node of node *b* is not a plateau node. Therefore, label of node *b* is equal to the value h_{21} contained in variable *label*. Node *b* is a plateau node. Therefore, the next node processed by algorithm 4 is node *c* of index $21+2^0 (=22)$ and of level 0. Index of left node of node *c* is $22-2^0 (=21)$. Left node of node *c* is a plateau node. Therefore, label of node *c* is equal to the hash of the first label extracted from the tail proof (value of node *c*) concatenated to the label of node *b* ($hash(h_{22}||h_{21})$). Node *c* is not a plateau node. Therefore, the next node processed by algorithm 4 is node *d* of the same index 22 and of level $0+1 (=1)$. Left node of node *d* is not a plateau node. Therefore, label of node *d* is equal to label of node *c* ($hash(h_{22}||h_{21})$). Node *d* is a plateau node. Therefore, the next node processed by algorithm 4 is node *s* of index $22+2^1 (=24)$ and of level 1. Index of left node of node *s* is $24-2^1 (=22)$. Left node of node *s* is a plateau node. Therefore, label of node *s* is equal to the hash of the second label extracted from the tail proof (label of node *r*) concatenated to the label of node *d*. Node *s* is not a plateau node. Therefore, the next node processed by algorithm 4 is node *t* of index 24 and of level 2. Index of left node of node *t* is $24-2^2 (=20)$. Left node of node *t* is a plateau node. Therefore, label of node *t* is equal to the hash of the label of node *s* concatenated to the first label extracted from the head proof (label of node *a*). Node *t* is not a plateau node. Therefore, the next node processed by algorithm 4 is node *u* of index 24 and of level 3. Left node of node *u* is not a plateau node. Therefore, label of node *u* is equal to label of node *t*. Node *u* is a plateau node. Therefore, the next node processed by algorithm 4 is the node of index $24+2^3 (=32)$ and level 3. Note that in figure 3, there is no node of index 32. In fact, everything works as if 32 was the index of the ending element. Consequently, the next node is node *y*. Left node of node *y* is node *u* which is a plateau node. Index of left node is $32-2^3 (=24)$. Therefore, the label of node *y* is equal to the hash of the third (and last) label extracted from the tail proof (label of node *x*) concatenated to the label of node *u*. Since node *y* is not a plateau

node, the next node processed by algorithm 4 is the node of index 32 and level 4 i.e. node z . Index of left node of node z is $32-2^4 (=16)$. Left node of node z is a plateau node. Therefore, label of node z is equal to the hash of the label of node y concatenated to the second label extracted from the head proof (label of node o). Since node z is not a plateau node, the next node processed by algorithm 4 is the node of index 32 and level 5 i.e. the node on top of node z i.e. the ending node. Index of left node of the ending node is $32-2^5 (=0)$. Left node of the ending node is a plateau node. Therefore, label of the ending node is equal to the hash of the label of node z concatenated to the last label extracted from the head proof (label of node q). Since there is no more labels to extract, neither from the tail proof nor from the head proof, Algorithm 4 compares the last computed label to the round token included in the timestamp. If the two labels are equal then the verification succeeds. If not, the verification fails.

4 Performances

We have implemented a prototype of our time-stamping scheme. We present the performances of our prototype in terms of space complexity. We focus on the number of hashing operations which are necessary to timestamp n documents (figure 4), and on the size of the timestamps (figure 5). In both figure 4 and figure 5, the X-axis stands for the number of requests. In figure 4, the Y-axis denotes the number of hashing operations made by the timestamping system whereas in figure 5, it denotes the number of digests included in the timestamps. From these two figures, we can see that the number of hashing operations is $O(n)$ and the number of digests included in

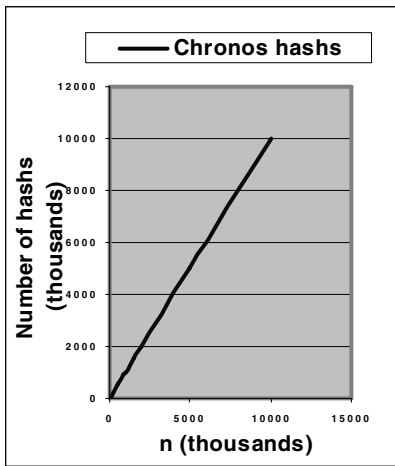


Fig. 4. Hashing cost

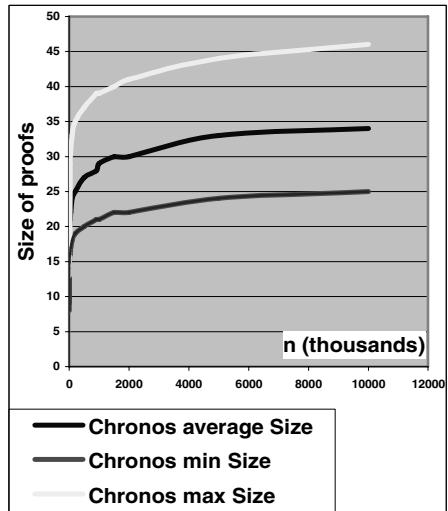


Fig. 5. Size of proofs

the timestamps is $O(\lg(n))$. In fact, our scheme has the same space complexity than the most efficient schemes used for timestamping (Merkle trees, threaded tree scheme, binary linking scheme...). However, compared to the binary linking scheme and to the threaded authentication tree scheme, our timestamping scheme has a smaller time complexity both for timestamping and verification. Indeed, our scheme needs as many concatenation operations as hashing operations, whereas binary linking scheme and threaded tree scheme need at least twice as many concatenation operations as hashing operations. Moreover, our scheme avoids the drawbacks of binary structures and accumulator systems.

Finally, let us mention that we could also compare our scheme to existing authenticated dictionary based on skip lists [10][14][15][16][17]. The reader can refer to [13] for such a comparison.

5 Conclusion

In this paper, we define a new totally ordered linking scheme based on skip lists. Our scheme offers better performances than existing totally ordered timestamping schemes. Moreover, it is easy to implement.

Our scheme is for a single server TSA. The main drawback of single server TSAs is that they are vulnerable to denials of service. In [18], we suggest some directions to implement a multi-server timestamping system. The main idea used in [18] is to randomly choose k servers among n . In a future work, we plan to develop a distributed version of our scheme based on skip lists, which would use this concept of k among n .

References

1. Bayer, D., Haber, S., Stornetta, W.: Improving the efficiency and reliability of digital time-stamping. In *Sequences'91: Methods in Communication, Security and Computer Science*, (1992) 329–334.
2. Benaloh, J., De Mare, M.: Efficient Broadcast time-stamping. Technical report 1, Clarkson University Department of Mathematics and Computer Science (1991).
3. Benaloh, J., De Mare, M.: One-way accumulators: A decentralized alternative to digital signatures. *Advances in Cryptology* (1993).
4. Buldas, A., Laud, P.: New Linking Schemes for Digital Time-Stamping. *First International Conference on Information Security and Cryptology* (1998).
5. Buldas, A., Laud, P., Lipmaa, A., Villemson J.: Time-stamping with Binary Linking Schemes. *Lecture Notes in Computer Science*, Vol. 1462. Springer-Verlag, Santa Barbara, USA (1998) 486–501.
6. Buldas, A., Lipmaa, A., Schoenmakers, B.: Optimally efficient accountable time-stamping. *Public Key Cryptography* (2000) 293–305.
7. Haber, S., Stornetta, W. S.: How to Time-stamp a Digital Document. *Journal of Cryptology: the Journal of the International Association for Cryptologic Research* 3(2) (1991).
8. Massias, H., Quisquater, J.J., Serret, X.: Timestamps : Main issues on their use and implementation. *Proc. of IEEE 8th International workshop on enabling technologies: Infrastructure for collaborative enterprises* (1999).

9. Massias, H., Quisquater, J.J., Serret, X.: Proc. of the 20th symposium on Information Theory in the Benelux (1999).
10. Maniatis, P., Giuli, T. J., Baker, M.: Enabling the long-term archival of signed documents through Time Stamping. Technical Report, Computer Science Department, Stanford University, California, USA, 2001.
11. Pugh, W.: Skip lists: a probabilistic alternative to balanced trees. *Communications of the ACM* (1990) 668–676.
12. Ansper, A., Buldas, A., Willemsen, J.: General linking schemes for digital time-stamping. Technical Report (1999).
13. Blibech, K., Gabillon, A.: Authenticated dictionary based on skip lists for timestamping systems. Proc. of the 12th ACM Conference on Computer Security, Secure Web Services Workshop (2005).
14. Maniatis, P., Baker, M.: Secure history preservation through timeline entanglement. Technical Report arXiv:cs.DC/0202005, Computer Science department, Stanford University, Stanford, CA, USA (2002).
15. Maniatis, P.: Historic Integrity in Distributed Systems. PhD thesis, Computer Science Department, Stanford University, Stanford, CA, USA (2003).
16. Goodrich, M., Tamassia, R.: Efficient authenticated dictionaries with skip lists and commutative hashing. Technical report, Johns Hopkins Information Security Institute (2000).
17. Goodrich, M., Tamassia, R., Schwerin, A.: Implementation of an authenticated dictionary with skip lists and commutative hashing (2001).
18. Bonnacaze, A., Liardet, P., Gabillon, A., Blibech, K.: A Distributed time stamping scheme. Proc. of the IEEE conference on Signal Image Technology and Internet based Systems (SITIS '05), Cameroon (2005).

A Semi-fragile Watermarking Scheme Based on SVD and VQ Techniques

Hsien-Chu Wu¹, Chuan-Po Yeh², and Chwei-Shyong Tsai³

¹ Department of Information Management, National Taichung Institute of Technology,
129 Sec. 3, San-min Road, Taichung, Taiwan 404, R.O.C.
wuhc@ntit.edu.tw

² Institute of Computer Science and Information Technology,
National Taichung Institute of Technology, 129 Sec. 3, San-min Road,
Taichung, Taiwan 404, R.O.C.
s18933108@ntit.edu.tw

³ Department of Management Information Systems, National Chung Hsing University,
250 Kuo Kuang Road, Taichung, Taiwan 402, R.O.C.
tsaics@nchu.edu.tw, tsaics@gmail.com

Abstract. Semi-fragile watermarking technique is effective for image authentication which is an important issue for safeguarding the integrity of digital images. It can be used to detect tampered regions and allow general image processing operation. In this paper, a novel semi-fragile watermarking technique based on SVD and VQ is proposed. The singular value can be against the image processing operation, and the coefficients of the matrix U and matrix V can represent the image feature. VQ is used to reduce the number of the image feature, and the VQ index is recorded to be the authentication criterion. The host image is transformed by SVD and image features are extracted from the SVD coefficients. VQ is then applied on these features where indices are obtained and embedded into the significant singular values. Experimental results show that the proposed scheme can locate the tampered regions and allow for JPEG lossy compression.

Keywords: Semi-fragile watermarking, Vector quantization (VQ), Singular value decomposition (SVD), Image authentication.

1 Introduction

Digital images are widely used in medical and military images. Therefore, for these images integrity and authenticity are important. Image authentication [1] can be used to provide such functions. It can be divided into two categories: digital signature techniques [2] and digital watermarking techniques [3-6]. The digital watermarking techniques for image authentication belong to fragile watermarking. The watermark is embedded into the protected image and then the watermarked image is transmitted to others. The receiver extracts the watermark and checks for tampering.

The fragile watermarking technique must have two properties [6]. First, it must detect whether an image is tampered or not. Second, it must locate where the tampered regions are. The easiest scheme is using LSB [7] in spatial domain. Wong and Memon proposed [3] a block-based watermarking scheme for image authentication in spatial domain. Each seven most bits of the pixel in the block is

input into a hash function. The output of the hash function is combined with the watermark by using the XOR operation, and then embedded into the image. This is an effective authentication scheme. However, this scheme will detect errors even if the meaning of the image does not change after modification. Digital image is often processed by general image processing operation, such as JPEG lossy compression, but this kind of scheme, like scheme [3], is so sensitive that the image can not allow any modification.

The semi-fragile watermarking technique is similar to the fragile watermarking, but it can allow some image processing operation. There are two requirements for the semi-fragile watermarking technique [8]. First, the malicious modification must be detected and the general image processing operation which does not change the significant meaning of the image must be allowed. Second, the malicious modification must be located. SVD and VQ are useful in the watermarking techniques. Many watermarking techniques [4-5, 9-12] are proposed based on them. Sun et al. [4] presented a semi-fragile watermarking technique based on SVD. The biggest singular value (SV) is quantified and then used to embed the watermark. This scheme is robust, but the variation of the block is ignored. If the original block is replaced by another block with SV similar to the original one, then it can not be detected. Lu et al. [5] proposed a VQ-based semi-fragile watermarking technique where the block is compared with the codewords for specific bit similar to the embedding bit. However, the robustness of this scheme is not good enough. For example, even if the quality factor of JPEG is set to be 100%, there will be a few detected errors.

In this paper, a hybrid watermarking scheme based on SVD and VQ is proposed. The SVD coefficients of U matrix and V matrix are processed by VQ to extract the image features, and these features are embedded into the biggest singular value of each block. According to the combination of SVD and VQ, experimental results show that the proposed scheme can locate the tampered regions meanwhile can allow for JPEG lossy compression.

The framework of this paper is organized as follows. Section 2 describes SVD and VQ individually. Section 3 describes our proposed scheme. The experimental results and conclusions are as shown in Section 4 and Section 5, respectively.

2 Review of SVD and VQ Techniques

2.1 Singular Value Decomposition

Let A be an $n \times n$ matrix. The SVD [4, 9-11] of A can be represented as follows:

$$\begin{aligned}
 A &= USV^T \\
 &= \begin{bmatrix} u_{0,0} & \cdots & u_{0,n-1} \\ \vdots & \ddots & \vdots \\ u_{n-1,0} & \cdots & u_{n-1,n-1} \end{bmatrix} \times \begin{bmatrix} s_0 & & \\ & \ddots & \\ & & s_{n-1} \end{bmatrix} \times \begin{bmatrix} v_{0,0} & \cdots & v_{0,n-1} \\ \vdots & \ddots & \vdots \\ v_{n-1,0} & \cdots & v_{n-1,n-1} \end{bmatrix}^T \\
 &= \sum_{i=0}^{n-1} s_i u_i v_i^T,
 \end{aligned} \tag{1}$$

where U and V are orthogonal matrices, and S is a diagonal matrix. u_i and v_i are the column vectors of matrix U and matrix V , respectively. Each s_i is the singular value (SV) of A , and $s_0 \geq s_1 \geq \dots \geq s_{n-1} \geq 0$. SVD has some properties for the digital images [9]. The luminance of the host image layer relates to SVs, and the intrinsic geometry properties of the image relates to the pair of u_i and v_i . The image quality will not decrease a lot after removing the smaller SVs, and this property can be used for lossy compression. The variations of SVs are slight after some general image processing manipulations and can be applied to the semi-fragile watermarking techniques.

2.2 Vector Quantization (VQ)

VQ is a lossy compression technique [5, 12-13]. The input is an image and a codebook. The image is first partitioned into blocks of the same size. Each block is computed for its Euclidean distances with the codewords in the codebook. Then the index of the codeword with the smallest Euclidean distance is output. Finally, an index table can be obtained, and this table represents the image. The compression rate is decided by block size and codebook size. For example, if the block size is 4×4 and codebook size is 256, the compression rate is 1/16.

3 The Proposed Scheme

The main idea behind the proposed scheme is to embed image features obtained from SVD coefficients, into the protected image. The block diagram of the proposed scheme is as shown in Figure 1.

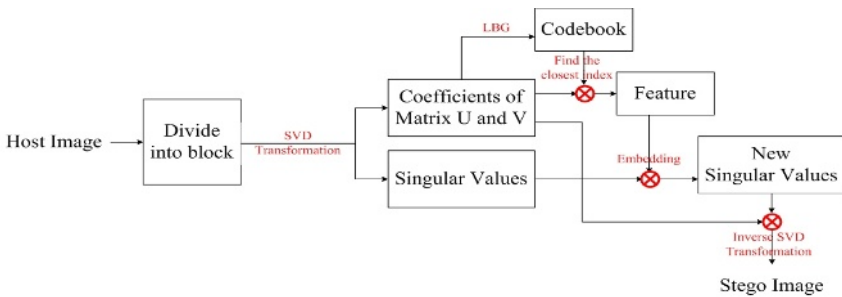


Fig. 1. The block diagram of the proposed scheme

3.1 Extracting Features

Figure 2 illustrates the flow for extracting the image features. Let I be the host image, and $I_{i,j}$ denote the block of host image I with index (i, j) . The size of block $I_{i,j}$ is $n \times n$. Each $I_{i,j}$ is divided into four sub-blocks $I_{i,j}(k, l)$, where $0 \leq k, l < 2$. Each sub-block $I_{i,j}(k, l)$ is computed by SVD transformation to get the SVD coefficients $C_{i,j}(k, l)$.

$$C_{i,j}(k, l) = SVD(I_{i,j}(k, l)) = \sum_{k=0}^{n/2-1} s_k u_k v_k^T, \tag{2}$$

where s_k is the singular value. u_k and v_k are column vectors. The first column vector u_0 and v_0 are used to extract feature of the block $I_{i,j}$. The feature $f_{k,l}$ of sub-block $I_{i,j}(k,l)$ is computed from the pair coefficients of u_0 and v_0 , and the feature $F(i,j)$ of the block $I_{i,j}$ is combined from $f_{k,l}$.

$$F(i,j) = (f_{0,0}, f_{0,1}, f_{1,0}, f_{1,1}). \tag{3}$$

$$f_{k,l} = (f_0, f_1, \dots, f_{n/2-1}). \tag{4}$$

$$f_x = \begin{cases} \frac{|u_{2x,0}| + |u_{2x+1,0}|}{2}, & \text{if } x < \frac{n}{4}; \\ \frac{|v_{2y,0}| + |v_{2y+1,0}|}{2}, & \text{if } x \geq \frac{n}{4}, \end{cases} \tag{5}$$

where $y = x \bmod n/2$. $u_{2x,0}$ and $v_{2y,0}$ are the elements of first column vector u_0 and v_0 , respectively. Note that, the purpose which takes the mean of the pair elements of the first column vector is to decrease the number of features in the block $I_{i,j}$. If there are more features, it will decrease the robustness. Finally, all of the features $F(i,j)$ are taken to train and generate a codebook C by using the most widely used Linde-Buzo-Gray (LBG) algorithm [13]. The size of the codebook C is 16.

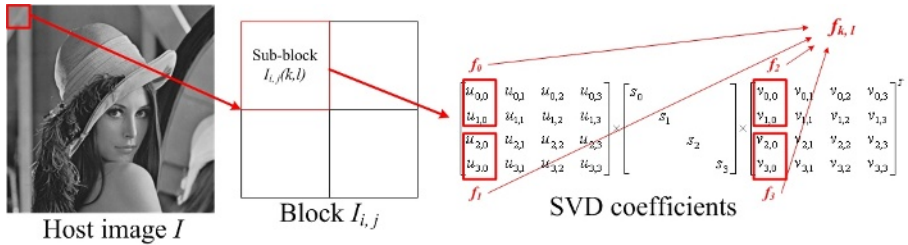


Fig. 2. Flow for extracting the features of the block $I_{i,j}$

3.2 The Embedding Process

In the embedding process, the features are embedded into the host image I . In Section 3.1, all of the features $F(i,j)$ are extracted from the host image I . These features $F(i,j)$ are further dealt with by VQ technique. Each feature $F(i,j)$ is compared with the codewords in the codebook C , and the index $ind_{i,j}$ of block $I_{i,j}$ can be obtained. The value of index $ind_{i,j}$ is from 0 to 15, and can be divided into four index bits $ind_{i,j}(x)$, where $0 \leq x < 4$ and each $ind_{i,j}(x)$ belongs to 0 or 1.

$$ind_{i,j}(x) = \frac{ind_{i,j}}{2^x}. \tag{6}$$

Each index bit $ind_{i,j}(x)$ is orderly embedded into the biggest SV s_0 of the sub-block $I_{i,j}(k,l)$. s_0 is quantified by quantification parameter Q , where Q is a power of 2. The

modified s'_0 of sub-block $I_{i,j}(k, l)$ is obtained by using LSB, and the equation is shown as follows.

$$s'_0 = \left(\left\lfloor \frac{s_0}{2Q} \right\rfloor \times 2 + ind_{i,j}(x) \right) \times Q + \frac{Q}{2}. \tag{7}$$

The modified sub-block $I'_{i,j}(k, l)$ is took after using inverse SVD transformation on modified SVs, original matrices U and V . The modified block $I'_{i,j}$ is generated if four sub-blocks are modified. Finally, the stego-image I' is obtained after each block is modified.

A specific signal also can be applied to our proposed scheme. If the specific signal is added into our scheme, this signal is taken as the watermark. Before embedding index bit into the s_0 , the new index bit $ind'_{i,j}(x)$ is obtained by using XOR operator on index bit $ind_{i,j}(x)$ and w , where w is a bit of the watermark and w belongs to 0 or 1.

$$ind'_{i,j}(x) = ind_{i,j}(x) \oplus w. \tag{8}$$

By using the same algorithm, the new index bit $ind'_{i,j}(x)$ can be embedded into SV s_0 . Therefore, the watermark can be embedded into host image. The size of the watermark is 4 times of the number of the blocks $I_{i,j}$. If there is no signal combined with index bit $ind_{i,j}(x)$, the index bit is taken as our watermark.

3.3 Extracting the Embedded Data

Let I'' be the received image, and block $I''_{i,j}$ is partitioned from I'' with index (i, j) . With the same process in Section 3.1, block $I''_{i,j}$ is divided into four sub-blocks $I''_{i,j}(k, l)$, and each sub-block $I''_{i,j}(k, l)$ can extract the feature $f''_{k,l}$. The feature $F''(i, j)$ of block $I''_{i,j}$ is combined from four features $f''_{k,l}$'s. By using the same codebook C , the index $ind''_{i,j}$ of block $I''_{i,j}$ can be obtained.

Each index $ind''_{i,j}$ of block $I''_{i,j}$ is compared with the extracted data $d_{i,j}$ to authenticate image. Here, the extracted data $d_{i,j}$ of block $I''_{i,j}$ is computed from the extracted bit $b_{i,j}(k, l)$ which is extracted from s_0 of the sub-block $I_{i,j}(k, l)$.

$$d_{i,j} = \sum_{k=0}^1 \sum_{j=0}^1 b_{i,j}(k, l) \times 2^{2k+l}. \tag{9}$$

$$b_{i,j}(k, l) = \left\lfloor \frac{s_0}{Q} \right\rfloor \bmod 2, \text{ where } s_0 \text{ is the SV of the sub-block } I_{i,j}(k, l). \tag{10}$$

The authentication process is to check whether the index $ind''_{i,j}$ of block $I''_{i,j}$ is the same as the extracted data $d_{i,j}$ of block $I''_{i,j}$ or not. If the index $ind''_{i,j}$ is the same as the extracted data $d_{i,j}$, the received image is not damaged, otherwise it is damaged.

3.4 Adjusting Robustness

Since the semi-fragile watermarking technique must be against some image processing operations, the codewords in the codebook C and the feature $f_{k,l}$ of sub-block $I_{i,j}(k, l)$ are adjusted to improve the robustness. Each feature $F(i, j)$ of block $I_{i,j}$

is classified by codewords. The original size of the codebook C is 16, and all of the features $F(i, j)$ are classified into 16 groups with the codewords in the centers, respectively. The distance between two codewords is computed, and the groups which belong to these two codewords are combined to be a bigger group if the distance is smaller than a threshold T . The size of the codebook is decreased, but the robustness is enhanced.

Since the feature $F(i, j)$ of block $I_{i, j}$ may change to another group after image processing operation, the feature $F(i, j)$ is adjusted to be close to the center of group (i.e. the codeword).

$$F'(i, j) = F(i, j) + (c_x - F(i, j)) \times p, \tag{11}$$

where $F'(i, j)$ is the feature of block $I_{i, j}$ after adjusting. c_x is the codeword with index x which the feature $F(i, j)$ belongs to. p is the parameter which controls the distance between the feature and the codeword. The value of p is between 0 and 1. The host image I is adjusted to enhance the robustness. Each $F(i, j)$ of block $I_{i, j}$ is changed to be $F'(i, j)$ by adjusting the column vector u_o and v_o of sub-block $I_{i, j}(k, l)$.

$$F'(i, j) = (f'_{0,0}, f'_{0,1}, f'_{1,0}, f'_{1,1}). \tag{12}$$

$$f'_{k,l} = (f'_0, f'_1, \dots, f'_{n/2-1}). \tag{13}$$

$$u'_{x,0} = \text{sign}(u_{x,0}) \times (|u_{x,0}| + (f'_y - f_y)), \text{ where } 0 \leq x < n/4, y = \lfloor x/2 \rfloor. \tag{14}$$

$$v'_{x,0} = \text{sign}(v_{x,0}) \times (|v_{x,0}| + (f'_y - f_y)), \text{ where } 0 \leq x < n/4, y = \lfloor x/2 \rfloor + n/4. \tag{15}$$

$\text{sign}(\cdot)$ is a function to indicate the number is positive or negative. If $\text{sign}(x) = 1$, the number x is positive. If $\text{sign}(x) = -1$, the number x is negative. After replacing $u_{x,0}$ and $v_{x,0}$ by $u'_{x,0}$ and $v'_{x,0}$, the modified column vector u'_o and v'_o of sub-block $I_{i, j}(k, l)$ can be obtained. The modified sub-block $I'_{i, j}(k, l)$ is obtained by using inverse SVD transformation with modified column vector u'_o , v'_o and original SV's. Finally, the modified image is obtained by adjusting each sub-block $I_{i, j}(k, l)$, and using this modified image to embed watermark which would be more robust.

4 Experimental Results and Discussions

SVD and VQ have intrinsic properties to be used on semi-fragile watermarking technique. The biggest SV represents the luminance of the block, and the U and V matrices relate the distribution of the edges. From the equation of SVD transformation (2), it is easy to observe that the biggest SV is multiplied by the first column vectors of matrix U and matrix V . Therefore, image layer 1 which is generated from the biggest SV and the first column vectors of matrix U and matrix V represents the image profile, and other layers represents the image variation, shown as Figure 3. Because the image layer 1 concentrates most energy in the block, the biggest SV and the first column vectors of matrix U and matrix V can be against some image processing operation. Figure 4 is an example which JPEG compression applies to the image block A. From Figure 3 and Figure 4, it is easy to observe that the biggest SV and the first column vectors of matrix U and matrix V is robust to be against JPEG

compression. The first column vectors of matrix U and matrix V are important enough to be used as the feature, and during the embedding process, modifying the SV will not change the U and V matrices so the feature will not be changed. However, it is difficult to embed the complete first column vectors into the image. VQ is used to solve this problem. The feature of the image block is processed by VQ, and the VQ index is obtained. This index can be used to represent the feature of image block, and it is embedded into the biggest SV of the image block. Furthermore, using VQ can provide the ability against the distortions of the feature.

$$\begin{aligned}
 A &= \begin{bmatrix} 132 & 210 & 205 & 194 \\ 156 & 213 & 201 & 189 \\ 158 & 208 & 201 & 177 \\ 182 & 206 & 196 & 177 \end{bmatrix} \\
 &= \begin{bmatrix} 0.50 & 0.71 & 0.13 & -0.48 \\ 0.51 & 0.09 & 0.41 & 0.75 \\ 0.50 & -0.10 & -0.85 & 0.15 \\ 0.50 & -0.69 & 0.30 & -0.42 \end{bmatrix} \begin{bmatrix} 755.50 & 0 & 0 & 0 \\ 0 & 38.11 & 0 & 0 \\ 0 & 0 & 6.16 & 0 \\ 0 & 0 & 0 & 2.53 \end{bmatrix} \begin{bmatrix} 0.42 & -0.89 & 0.15 & -0.11 \\ 0.55 & 0.13 & -0.19 & 0.80 \\ 0.53 & 0.21 & -0.61 & -0.55 \\ 0.49 & 0.38 & 0.75 & -0.22 \end{bmatrix}^T \\
 &= \begin{bmatrix} 155.84 & 207.60 & 199.14 & 182.79 \\ 158.78 & 211.52 & 202.90 & 186.24 \\ 155.54 & 207.20 & 198.76 & 182.44 \\ 158.13 & 210.64 & 202.06 & 185.47 \end{bmatrix} \begin{bmatrix} -24.09 & 3.53 & 5.66 & 10.36 \\ -2.96 & 0.43 & 0.70 & 1.28 \\ 3.29 & -0.48 & -0.77 & -1.42 \\ 23.48 & -3.44 & -5.52 & -10.10 \end{bmatrix} \\
 &\quad \text{Image layer 1} \qquad \qquad \qquad \text{Image layer 2} \\
 &+ \begin{bmatrix} 0.12 & -0.14 & -0.47 & 0.58 \\ 0.39 & -0.48 & -1.56 & 1.91 \\ -0.80 & 0.99 & 3.21 & -3.94 \\ 0.28 & -0.35 & -1.13 & 1.39 \end{bmatrix} \begin{bmatrix} 0.13 & -0.98 & 0.67 & 0.27 \\ -0.20 & 1.53 & -1.04 & -0.43 \\ -0.04 & 0.29 & -0.20 & -0.08 \\ 0.11 & -0.86 & 0.58 & 0.24 \end{bmatrix} \\
 &\quad \text{Image layer 3} \qquad \qquad \qquad \text{Image layer 4}
 \end{aligned}$$

Fig. 3. An example of SVD transformation where A is a image block

$$\begin{aligned}
 A_i &= \begin{bmatrix} 132 & 205 & 211 & 198 \\ 148 & 210 & 203 & 189 \\ 165 & 209 & 194 & 180 \\ 177 & 210 & 192 & 178 \end{bmatrix} \\
 &= \begin{bmatrix} 0.50 & 0.71 & 0.49 & -0.12 \\ 0.50 & 0.21 & -0.70 & 0.47 \\ 0.50 & -0.31 & -0.25 & -0.77 \\ 0.50 & -0.60 & 0.46 & 0.42 \end{bmatrix} \begin{bmatrix} 754.57 & 0 & 0 & 0 \\ 0 & 40.55 & 0 & 0 \\ 0 & 0 & 3.11 & 0 \\ 0 & 0 & 0 & 0.02 \end{bmatrix} \begin{bmatrix} 0.41 & -0.84 & 0.35 & 0.02 \\ 0.55 & -0.08 & -0.82 & -0.11 \\ 0.53 & 0.37 & 0.22 & 0.73 \\ 0.49 & 0.39 & 0.39 & -0.68 \end{bmatrix}^T
 \end{aligned}$$

Fig. 4. An example of SVD transformation where A_i is the image block after JPEG compression

In our experiment, a 512×512 grayscale image is the host image. The size of block is 8×8 , and the quantification parameter $Q = 32$. Figure 5(a) is the stego-image of the image “Baboon” which the watermark is not embedded into. The parameter p is set to be 0.75 and the PSNR of the stego image is 27.80 dB. Figure 5(b) is the stego-image

of the image “Lena” with the embedded watermark. The parameter p is set to be 0.8 and the PSNR of the stego image is 30.40 dB. For the lossy compression, both of the stego-images “Baboon” and “Lena” have no error after JPEG compression with quality factor 75%.

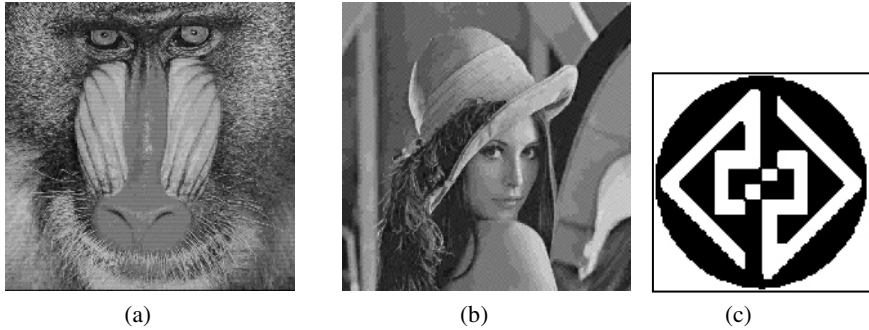


Fig. 5. (a) Stego-image “Baboon”, (b) stego-image “Lena” and (c) 128×128 binary watermark

In the experiment, each region which is damaged by malicious attack can be identified. Figure 6(a) shows the stego-image attacked by drawing black lines. Figure 6(b) is the authentication result, and the damaged region is located by the darker block. Figure 7 shows the attack which the right eye of the stego-image is replaced by the left eye.

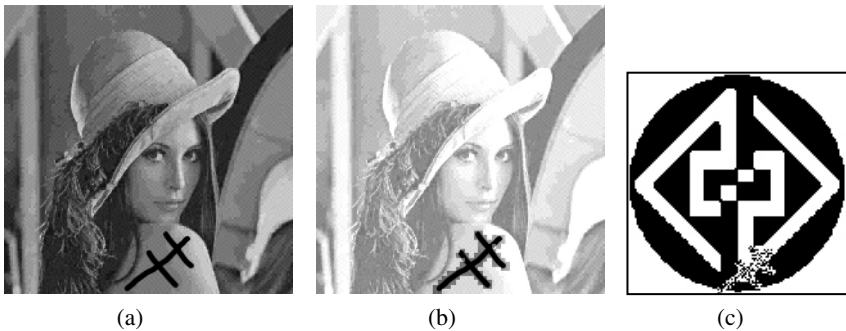


Fig. 6. (a) The attacked stego-image, (b) authentication result and (c) the extracted watermark

The hybrid watermarking scheme based on SVD and VQ is more suitable for the semi-fragile watermarking technique. In [4], the semi-fragile watermarking scheme only uses SVD. The scheme is concentrated only on the biggest SV, but not concerned with the edge in the block. If the image block is replaced by another one which has the same SVs but the matrix U and V are different, this scheme can not detect the change. Take Figure 8 as an example, the image blocks in Figure 8(a) and Figure 8(b) have the similar biggest SV with A in Figure 3, but the first column vectors are different. If the image block A in Figure 3 is replaced by the image block in Figure 8, the semi-fragile

watermarking scheme can not detect the change. Our proposed scheme uses the first column vector of matrix U and matrix V to be the feature so the variation of the block can be identified. Compared with [5], the semi-fragile watermarking scheme only uses VQ, and it is sensitive to detect the change. However, the embedded watermark in [5] still will be tampered even if the quality factor of JPEG is set to be 100%, but our proposed scheme can allow JPEG compression. The optimal semi-fragile watermarking scheme must satisfy two properties. One is to detect the tampered regions accurately and the other is to have the ability of allowing general image processing operation. Both [4] and [5] have good performance to one property but are weak to another property. Our proposed scheme is based on SVD and VQ, and it has good performance to these two properties.

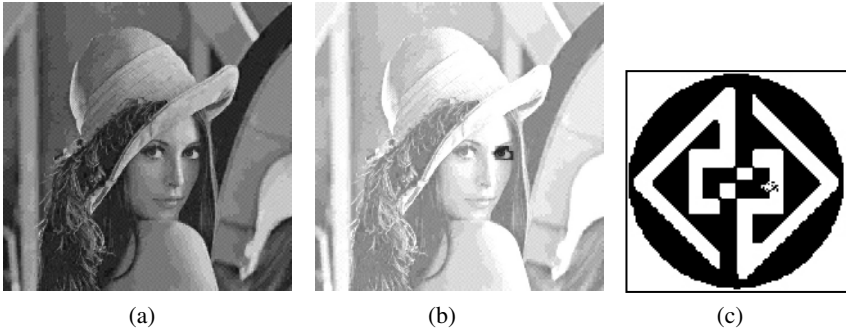


Fig. 7. (a) The attacked stego-image, (b) authentication result and (c) the extracted watermark

$$\begin{aligned}
 A_2 &= \begin{bmatrix} 132 & 156 & 158 & 182 \\ 210 & 213 & 208 & 206 \\ 205 & 201 & 201 & 196 \\ 194 & 189 & 177 & 177 \end{bmatrix} \\
 &= \begin{bmatrix} 0.42 & 0.89 & -0.15 & -0.11 \\ 0.55 & -0.13 & 0.19 & 0.80 \\ 0.53 & -0.21 & 0.61 & -0.55 \\ 0.49 & -0.38 & -0.75 & -0.22 \end{bmatrix} \begin{bmatrix} 755.50 & 0 & 0 & 0 \\ 0 & 38.11 & 0 & 0 \\ 0 & 0 & 6.16 & 0 \\ 0 & 0 & 0 & 2.53 \end{bmatrix} \begin{bmatrix} 0.50 & -0.71 & -0.13 & -0.48 \\ 0.51 & -0.09 & -0.41 & 0.75 \\ 0.50 & 0.10 & 0.85 & 0.15 \\ 0.50 & 0.69 & -0.30 & -0.42 \end{bmatrix} \\
 &\quad (a) \\
 A_3 &= \begin{bmatrix} 205 & 205 & 124 & 135 \\ 252 & 124 & 254 & 205 \\ 254 & 154 & 203 & 148 \\ 155 & 154 & 254 & 153 \end{bmatrix} \\
 &= \begin{bmatrix} 0.44 & -0.79 & -0.25 & 0.34 \\ 0.56 & 0.39 & 0.49 & 0.54 \\ 0.51 & -0.16 & 0.37 & -0.76 \\ 0.48 & 0.44 & -0.75 & -0.13 \end{bmatrix} \begin{bmatrix} 755.58 & 0 & 0 & 0 \\ 0 & 114.34 & 0 & 0 \\ 0 & 0 & 73.57 & 0 \\ 0 & 0 & 0 & 26.50 \end{bmatrix} \begin{bmatrix} 0.58 & -0.33 & 0.68 & -0.30 \\ 0.41 & -0.63 & -0.66 & -0.03 \\ 0.56 & 0.69 & -0.29 & -0.35 \\ 0.43 & 0.14 & 0.10 & 0.89 \end{bmatrix} \\
 &\quad (b)
 \end{aligned}$$

Fig. 8. (a) An example of SVD transformation where A_2 is the matrix transpose of A in Figure 5 (b) An example of SVD transformation where A_3 is the image block with the biggest SV close to A 's biggest SV

By adjusting the parameter p and quantification parameter Q , the robustness of the watermark and image quality can be adjusted. Take the image "Lena" as an example, if the parameter p is set to be 0.22 and quantification parameter Q is set to be 4, the PSNR is 44.52 dB, but the ability of allowing image processing operation is weak. Note that, if the parameter p is smaller than 0.22 or quantification parameter Q is smaller than 4, there will be some errors because of converting real numbers into integers.

5 Conclusions

In this paper, a semi-fragile watermarking technique is proposed. The relationship between the image and its SVD coefficients is utilized. The first column vectors of matrix U and matrix V are used as the image features, and these features are processed by VQ to increase robustness. The output of VQ is finally embedded into the biggest SV. The relation between robustness and image quality can be adjusted by the parameter p and quantification parameter Q . Whether the watermark will be added to the feature or not, will be decided by user. From experimental results, our proposed scheme can resist the JPEG lossy compression. By applying the proposed semi-fragile watermarking technique, the tampered regions can be detected and located effectively.

References

1. Lu, C.S., Liao, H.Y.M.: Structural digital signature for image authentication an incidental distortion resistant scheme. *IEEE Transactions on Multimedia*, Vol. 5, June 2003, pp. 161-173
2. Lou, D.C., Liu J.L.: Fault resilient and compression tolerant digital signature for image authentication. *IEEE Transactions on Consumer Electronics*, Vol. 46, Feb. 2000, pp. 31-39
3. Wong, P., Memon, N.: Secret and public key image watermarking schemes for image authentication and ownership verification. *IEEE Transactions on Image Processing*, Vol.10, Oct. 2001, pp. 1593-1601
4. Sun, R., Sun, H., Yao, T.: A SVD and quantization based semi-fragile watermarking technique for image authentication. 2002 6th International Conference on Signal Processing, Vol. 2, Aug. 2002, pp. 1592-1595
5. Lu, Z.M., Liu, C.H., Xu, D.G., Sun, S.H.: Semi-fragile image watermarking method based on index constrained vector quantization. *Electronics Letters*, Vol. 39, Jan. 2003, pp. 35-36
6. Li, C.T.: Digital fragile watermarking scheme for authentication of JPEG images. *IEE Proc.-Vis. Image Signal Process*, Vol. 151, Dec. 2004, pp. 460-466
7. Bender, W., Gruhl, D., Morimoto, N., Lu, A.: Techniques for data hiding. *IBM System Journal*, Vol. 35, No. 3-4, 1996, pp. 313-337
8. Wu, Y.T., Shih, F.Y.: An adjusted-purpose digital watermarking technique. *Pattern Recognition*, Vol.37, 2004, pp. 2349 – 2359
9. Bao, P., Ma, X.: Image adaptive watermarking using wavelet domain singular value decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.15, Jan. 2005, pp.96-102
10. Chang, C.C., Tsai, P., Lin, C.C.: SVD-based digital image watermarking scheme. *Pattern Recognition Letters*, Vol. 26, July 2005, pp. 1577-1586
11. Huang, F.J., Guan, Z.H.: A hybrid SVD-DCT watermarking method based on LPSNR. *Pattern Recognition Letters*, Vol. 25, Nov. 2004, pp. 1769-1775
12. JeHwang, J.: Digital image watermarking employing codebook in vector quantization. *Electronics Letters*, Vol. 39, May 2003, pp. 840-841
13. Linde, Y., Buzo, A., Gray, R.: An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications*, Vol. 28, Jan. 1980, pp.84-95

New Constructions of Universal Hash Functions Based on Function Sums

Khoongming Khoo¹ and Swee-Huay Heng²

¹ DSO National Laboratories,
20 Science Park Drive, Singapore 118230
kkhoongm@dso.org.sg

² Faculty of Information Science and Technology, Multimedia University,
Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia
shheng@mmu.edu.my

Abstract. In this paper, we propose a generalization of the SQUARE hash function family to the function sum hash, which is based on functions with low maximal differential over arbitrary Abelian groups. These new variants allow the designer to construct SQUARE-like hash functions on different platforms for efficient and secure message authentication. A variant using functions with low algebraic degree over a finite field is also proposed which enables the user to use a shorter key. For more versatility, we also propose a trade-off between the hash key length and security bound. Finally, we show that we can use an SPN structure in the function sum hash to construct a provably secure MAC with performance which is several times faster than the traditional CBC-MAC. Moreover, there are implementation advantages like parallelizability to increase the speed further and re-use of cipher components which help save on implementation resources.

Keywords: Message authentication codes, universal hash functions, low maximal differential, low algebraic degree, substitution permutation network (SPN).

1 Introduction

Message authentication codes (MAC) are important cryptographic functions which provide data integrity and data origin authentication. In 1981, Wegman and Carter proved that we can construct a secure MAC by encrypting a family of universal hash functions [17]. Since then there have been many papers on efficient construction of universal hash functions for secure MAC [13, 14, 15, 10, 8, 7, 11, 12]. Some of these papers like [8, 7, 11] focus on software implementations of universal hash over the field \mathbf{Z}_p . There were also some papers for universal hash functions on hardware implementations. For example, Krawczyk proposed universal hash functions that are linear with respect to bit-wise exclusive-or (XOR) which is based on polynomials and Linear Feedback Shift Registers (LFSRs) [13, 14]. Shoup [15] studied efficient software implementations of the above primitive. Universal hash functions from exponential sums

over finite fields and Galois rings which are suitable for both hardware and software implementations were proposed in [10].

The SQUARE hash functions [7] are based on \mathbf{Z}_p -multiplication, and they provide efficient, secure MAC's on certain platforms (e.g. a Pentium processor) which enable fast modular multiplication. In Section 3, we propose a generalization, called the function sum hash (FSH), by replacing the squaring operation $f(x) = x^2 \pmod{p}$ in the SQUARE hash with a function having low maximal differential in an arbitrary Abelian group. Some possible candidates for replacement include the perfect nonlinear functions over $GF(p^n)$ for $p \neq 2$ and the almost perfect nonlinear functions over $GF(2^n)$.

In [11], a variant of the SQUARE hash which uses a short fixed length key is proposed. It was shown in [11] that this gives a hash function with comparable security but avoided the use of a LFSR (which does not have provable security) to generate a long key. Moreover, the short key needs less storage space which is suitable for applications with memory constraints. In Section 3, we provide a generalization of this concept: to construct a short-key function sum hash, which we call FSH'. We show that if the underlying function has low algebraic degree in a finite field F , then the security is of comparable magnitude to FSH. We also extend the FSH' to a more flexible setting where we allow the key to be r times longer in order to obtain a r -fold increase in security. Thus there is a storage-security trade-off.

In Section 4 we consider some applications of our generalized construction. When applied to the function x^2 in the finite field \mathbf{Z}_p , we obtain the square hash SQH and SQH' of [7, 11]. These are very efficient when implemented on Pentium processors which can perform fast squaring. However, when applied on hardware, the square hash is not suitable because modular multiplication is slow. In that case, we may apply the FSH construction to the Gold function x^{2^i+1} and the FSH' construction to the cubic function x^3 over $GF(2^n)$. They are efficient on dedicated hardware which performs efficient finite field multiplications based on optimal normal basis [1]. In SQH and SQH', k squarings are needed to process k message blocks while in our $GF(2^n)$ construction, k fast multiplications are needed to process k message blocks.

Since we have shown that the security of the FSH depends on the differential property of the nonlinear function, it is tempting to use the many well-known nonlinear S-boxes used in block ciphers which have optimally small maximal differential. However, for adequate security, these S-boxes will have to be sufficiently large. For example, to get a security bound in the magnitude of 2^{-64} , we need to use 64-bit S-boxes. But to implement it as a look-up-table requires $2^{64} \times 64 = 2^{72}$ memory which is not feasible in practice. One solution is to use efficient computation like the function x^{2^i+1} over $GF(2^n)$ which only needs one finite field multiplication which is efficient on optimized hardware platforms. The other is to use small look-up-tables in a substitution permutation network (SPN) as in block ciphers like AES and Square. In Section 5, we show how to construct fast and secure function sum hash from a combination of small S-boxes and Maximum Distance Separable (MDS) linear transform. We also

explain how this construction uses similar components as a CBC-MAC but gives rise to a provably secure MAC which is more efficient.

2 Preliminaries

A measure of the efficiency of block ciphers against differential cryptanalysis is to possess low maximum differential. Let G be an Abelian group and $f : G \rightarrow G$. The *maximum differential* of f is:

$$\Delta_f = \max_{a \neq 0, b \in G} |\{x \in G \mid f(x+a) - f(x) = b\}|.$$

Besides protecting against differential cryptanalysis, we show in this paper that functions with low maximum differential can be used to construct good universal hash functions and secure MAC's.

Let H be a family of functions from a domain D to a range R . The probabilities below, denoted by $\Pr_{h \in H}[\cdot]$, are taken over the choice of $h \in H$. Let R be a finite Abelian group and denote by $'-'$ the group subtraction operation. H is a Δ -universal family of hash functions if for all $x, y \in D$ with $x \neq y$ and all $a \in R$, $\Pr_{h \in H}[h(x) - h(y) = a] = \frac{1}{|R|}$. H is an ϵ -almost- Δ -universal (ϵ - Δ U) family of hash functions if $\Pr_{h \in H}[h(x) - h(y) = a] \leq \epsilon$.

We have the following collision bounds for universal hash functions:

Proposition 1. [16] *If H is an ϵ - Δ U family of hash functions, then $\epsilon \geq 1/|R|$.*

Universal hash functions can be used to construct *Message Authentication Codes* (MAC) via the Wegman-Carter approach [17]. The MAC tag is given by the value $h(m)$ exclusive-or-ed with the one-time-pad *OTP* as follows:

$$MAC_{h,OTP}(m) = h(m) \oplus OTP$$

where h is a randomly chosen hash function from the family H and *OTP* is a random one-time-pad. We note that the communicating parties must share the secret key pair (h, OTP) in this scenario. However, it is not practical to generate one-time-pads long enough to handle long messages. In [4], Brassard proposed that we substitute the one-time-pad encryption with a computationally secure encryption scheme, for example AES.

In this paper, we denote the characteristic of a finite field F by $char(F)$ and the algebraic degree of a function $f : F \rightarrow F$ by $deg(f)$.

3 Universal Hash Functions Based on Function Sums

We introduce the *function sum hash* (FSH) family which can be considered as a generalization of the square hash (SQH) by Etzel *et al* [7]. First, let us review the SQH hash family:

Definition 1. Let $k > 0$ be an integer and p be an odd prime. Let $x = (x_1, \dots, x_k)$ and $m = (m_1, \dots, m_k)$ where $x_i, m_i \in \mathbf{Z}_p$. The SQH family of functions is defined as follows: $SQH \triangleq \{h_x : \mathbf{Z}_p^k \rightarrow \mathbf{Z}_p \mid x \in \mathbf{Z}_p^k\}$ where $h_x(m) \triangleq \sum_{i=1}^k (m_i + x_i)^2 \pmod{p}$.

Theorem 1. [7] SQH is a Δ -universal family of hash functions.

In FSH, we substitute the function $x^2 \pmod{p}$ in SQH by some function $f : G \rightarrow G$ in an arbitrary group G . Then we show that FSH will give good hash families when the underlying function f has low maximum differential.

Definition 2. Let G be a finite Abelian group, $f : G \rightarrow G$ and $k > 0$ be an integer. Let $x = (x_1, \dots, x_k)$ and $m = (m_1, \dots, m_k)$ where $x_i, m_i \in G$. The FSH family of functions is defined as follows: $FSH \triangleq \{h_x : G^k \rightarrow G \mid x \in G^k\}$ where $h_x(m) \triangleq \sum_{i=1}^k f(m_i + x_i)$.

Theorem 2. In definition 2, let $f : G \rightarrow G$ have maximal differential Δ_f . Then FSH is an ϵ - $A\Delta$ universal family of hash functions with $\epsilon \leq \frac{\Delta_f}{|G|}$.

Proof. Fix some $\delta \in G$ and let $m \neq n$ be two messages from G^k . Since $m \neq n$, there is some i for which $m_i \neq n_i$. Without loss of generality, we assume $m_1 \neq n_1$. Then, for any choice of x_2, \dots, x_k , we have:

$$\begin{aligned} \Pr_{x \in G^k} [h_x(m) - h_x(n) = \delta] &= \Pr_{x \in G^k} \left[\sum_{i=1}^k (f(m_i + x_i) - f(n_i + x_i)) = \delta \right] \\ &= \Pr_{x_1 \in G} [f(m_1 + x_1) - f(n_1 + x_1) = C], C = \delta - \sum_{i=2}^k (f(m_i + x_i) - f(n_i + x_i)) \\ &= \Pr_{y \in G} [f(y) - f(n_1 - m_1 + y) = C] \leq \frac{\Delta_f}{|G|}. \quad \square \end{aligned}$$

Recently, Heng and Kurosawa [11] introduced SQH', a variant of the square hash SQH, which uses only one short scalar in their key generation instead of k independent scalars.

Definition 3. Let $k > 0$ be an integer and p be an odd prime. Let $x \in \mathbf{Z}_p$ and $m = (m_1, \dots, m_k)$ where $m_i \in \mathbf{Z}_p$. The SQH' family of functions is defined as follows: $SQH' \triangleq \{h_{x_1} : \mathbf{Z}_p^k \rightarrow \mathbf{Z}_p \mid x_1 \in \mathbf{Z}_p\}$ where $h_{x_1}(m) \triangleq \sum_{i=1}^k (m_i + x_1^i)^2 \pmod{p}$.

Theorem 3. [11] SQH' is a ϵ - $A\Delta$ -universal family of hash functions with $\epsilon \leq \frac{k}{p}$.

Thus, one can still generate a secure MAC from SQH' without relying on the pseudo random number generator (PRNG) which has to generate the key x from a random seed. This is an advantage as we do not have to rely anymore on the PRNG, whose security we have to assume in the security proof. We can also have a similar variant for FSH, which we call FSH'.

Definition 4. Let F be a finite field and $k > 0$ be an integer. Let $x_1 \in F$ and $m = (m_1, \dots, m_k)$ where $m_i \in F$. The FSH' family of functions is defined as follows: $FSH' \triangleq \{h_{x_1} : F^k \rightarrow F | x_1 \in F\}$ where $h_{x_1}(m) \triangleq \sum_{i=1}^k f(m_i + x_1^i)$.

Theorem 4. In definition 4, let $\text{deg}(f) = d$ and $\text{char}(F) \nmid d$. Then FSH' is an ϵ - $A\Delta$ universal family of hash functions with $\epsilon \leq \frac{(d-1)k}{|F|}$.

Proof. Since $\text{deg}(f) = d$, we see that for all i ,

$$\begin{aligned} & \text{deg}(f(m_i + x_1) - f(n_i + x_1)) \leq d - 1 \\ \implies & \text{deg}(f(m_i + x_1^i) - f(n_i + x_1^i)) \leq i(d - 1) \\ \implies & \text{deg}\left(\sum_{i=1}^k (f(m_i + x_1^i) - f(n_i + x_1^i))\right) \leq k(d - 1). \end{aligned}$$

Fix some $\delta \in F$ and let $m \neq n$ be two messages from F^k .

$$\begin{aligned} & \Pr_{x_1 \in F} [h_{x_1}(m) - h_{x_1}(n) = \delta] \\ = & \Pr_{x_1 \in F} \left[\sum_{i=1}^k f(m_i + x_1^i) - f(n_i + x_1^i) = \delta \right] \leq \frac{k(d - 1)}{|F|}. \end{aligned}$$

This is because an equation of degree $(d - 1)k$ can have at most $(d - 1)k$ roots in the field F . □

As suggested by Heng and Kurosawa [11], there exists trade-off between the key length and the security bound of the hash function. Precisely, we can have a r -fold increase in security when we increase the key length by r times. We show in the following theorem that this applies in the general case. We define a function FSH'_r which expands the key of FSH' by r times.

Theorem 5. Let $0 < r < k$ be integers, F be a finite field where $\text{char}(F) \nmid d$ and $f : F \rightarrow F$ satisfies $\text{deg}(f) = d$. Let $s = k \bmod r$. Let m be a k -block message indexed by

$$\begin{aligned} m = & (m_{1,1}, \dots, m_{1,\lceil k/r \rceil} \dots m_{s,1}, \dots, m_{s,\lceil k/r \rceil}, \\ & m_{s+1,1}, \dots, m_{s+1,\lceil k/r \rceil} \dots m_{r,1}, \dots, m_{r,\lceil k/r \rceil}) \end{aligned}$$

and $x = (x_1, x_2, \dots, x_r)$ be a r -block key where $m_{i,j}, x_i \in F$. Then the FSH'_r hash defined by

$$h_x(m) \triangleq \sum_{i=1}^s \sum_{j=1}^{\lceil k/r \rceil} f(m_{i,j} + x_i^j) + \sum_{i=s+1}^r \sum_{j=1}^{\lceil k/r \rceil} f(m_{i,j} + x_i^j)$$

is ϵ - $A\Delta$ universal with $\epsilon \leq \frac{(d-1)\lceil k/r \rceil}{|F|}$.

Proof. Fix some $\delta \in F$ and let $m \neq n$ be two messages from F^k . Let C_1 be the number of distinct indices $1 \leq i \leq s$ such that $m_{i,j} \neq n_{i,j}$ for some j . Let C_2 be the number of distinct indices $s + 1 \leq i \leq r$ such that $m_{i,j} \neq n_{i,j}$ for some j . Note that $m \neq n$ implies $C_1 + C_2 \geq 1$.

$$\begin{aligned} & \Pr_{x \in F^r} [h_x(m) - h_x(n) = \delta] \\ &= \Pr_{x \in F^r} \left[\sum_{i \in C_1} \sum_{j=1}^{\lfloor k/r \rfloor} (f(m_{i,j} + x_i^j) - f(n_{i,j} + x_i^j)) \right. \\ & \quad \left. + \sum_{i \in C_2} \sum_{j=1}^{\lfloor k/r \rfloor} (f(m_{i,j} + x_i^j) - f(n_{i,j} + x_i^j)) = \delta \right] \\ &\leq \frac{((d-1)\lfloor k/r \rfloor)^{C_1} \times ((d-1)\lfloor k/r \rfloor)^{C_2}}{|F|^{C_1+C_2}} \leq \frac{(d-1)\lfloor k/r \rfloor}{|F|}. \end{aligned}$$

As shown in the proof of Theorem 4, the first group of summands has degree at most $(d-1)\lfloor k/r \rfloor$ and the second group of summands has degree at most $(d-1)\lfloor k/r \rfloor$. We upper bound the number of solution tuples $(x_i, i \in C_1 \text{ or } C_2)$ by the product of the number of solutions for each x_i when the other variables $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_r)$ are known. It is clear that our upper bound is maximal when $C_1 = 1$ and $C_2 = 0$. □

4 Applications to Finite Fields and Analogues for the SQUARE Hash over $GF(2^n)$

A function $f(x)$ with the best possible differential $\Delta_f = 1$ is called a *perfect nonlinear function*. By Theorem 2, applying perfect nonlinear functions in the FSH construction gives Δ -universal hash function (lowest $\epsilon = \frac{1}{|G|}$). We consider the case where $G = F$ is a finite field in Definition 2. A perfect nonlinear function exists in a finite field $GF(p^n)$ only when $p \neq 2$. This is because in $GF(2^n)$, if x satisfies $f(x) + f(x + a) = b$, then $x + a$ also satisfies the equation. The known perfect nonlinear functions are of the form x^d which are listed in Table 1. Corollary 1 is obtained by applying Theorem 2 to perfect nonlinear functions.

Corollary 1. *In definition 2, let $G = GF(p^n)$ be a finite field where $p \neq 2$ and $f : GF(p^n) \rightarrow GF(p^n)$ be a perfect nonlinear function (as defined in Table 1). Then FSH gives a Δ -universal hash family.*

The efficient and secure SQUARE hash is a special case with $f(x) = x^2$ defined on \mathbf{Z}_p .

How about applying the FSH construction to the finite field $GF(2^n)$? The functions with the best differential properties are the *almost perfect nonlinear* (APN) functions $f : GF(2^n) \rightarrow GF(2^n)$ which has maximal differential $\Delta_f = 2$. In Table 2, we list the known APN functions in the literature (extracted from [5]). Corollary 2 is obtained by applying Theorem 2 to APN functions.

Table 1. $f(x)$ is perfect nonlinear on $GF(p^n)$, $p \neq 2$ [9]

Function $f(x)$	Condition
$ax^2 + bx + c$	$a \neq 0$
x^{p^k+1}	$n/\gcd(n, k)$ odd
$x^{(3^k+1)/2}$	$p = 3, k$ odd and $\gcd(k, n) = 1$

Table 2. Exponents r such that x^r is APN on $GF(2^n)$ [5]

Exponent r	Condition
$2^j + 1$ (Gold)	$\gcd(j, n) = 1$
-1 (Inverse)	n odd
$2^{2j} - 2^j + 1$ (Kasami)	$\gcd(j, n) = 1$
$2^{(n-1)/2} + 3$ (Welch)	n odd
$2^{2j} + 2^j - 1$ (Niho)	$4j + 1 \equiv 0 \pmod{n}, n$ odd
$2^{4j} + 2^{3j} + 2^{2j} + 2^j - 1$ (Dobbertin)	$n = 5j, n$ odd

Corollary 2. *In definition 2, let $G = GF(2^n)$ and let $f : GF(2^n) \rightarrow GF(2^n)$, $f(x) = x^r$, be an APN function (r as defined in Table 2). Then FSH gives an ϵ - $A\Delta$ universal family of hash functions with $\epsilon \leq \frac{1}{2^{n-1}}$.*

On certain platforms which allow fast $GF(2^n)$ computations, FSH based on APN functions may give fast and secure hash functions. For example, consider the following function defined on $GF(2^n)$:

$$h_x(m) = \sum_{i=1}^k (m_i + x_i)^{2^j+1}, \quad \gcd(j, n) = 1, \tag{1}$$

It is an ϵ - $A\Delta$ universal family of hash functions with $\epsilon \leq \frac{1}{2^{n-1}}$ by Corollary 2. We can write x^{2^j+1} as $x^{2^j} \cdot x$ which can be computed very efficiently on cryptographic platforms which use an optimal normal basis (ONB) for $GF(2^n)$ computations [1]. This is because squaring, which is equivalent to cyclic shifts in normal bases, is free and we perform just k optimized $GF(2^n)$ -multiplications to compute $h_x(m)$. This is analogous to a software implementation of SQH which uses k squarings modulo p . Moreover, they have comparable security bounds, the bound for SQH is $1/|R|$, $|R| = p$ while that for this construction is $2/|R|$, $R = 2^n$.

Corollary 3 is a direct consequence of applying Theorem 4 on quadratic functions.

Corollary 3. *In definition 4, let $\text{char}(F) \neq 2$ and $f(x) = ax^2 + bx + c$ where $a, b, c \in F$. Then FSH' gives an ϵ - $A\Delta$ universal family of hash functions with $\epsilon \leq \frac{k}{|F|}$.*

The SQH' hash function proposed in [11] is a special case with $f(x) = x^2$ on \mathbf{Z}_p . However in $GF(2^n)$, we need to use low degree functions in the FSH' construction which are not quadratic.

Corollary 4. *In definition 4, let $f : GF(2^n) \rightarrow GF(2^n)$ be defined by $f(x) = x^3$ and n be odd. Then FSH' gives an ϵ - $A\Delta$ universal family of hash functions with $\epsilon \leq \frac{k}{2^{n-1}}$.*

As in the analysis of equation (1), the FSH' construction based on x^3 needs k optimized $GF(2^n)$ -multiplications which is analogous to the k squarings in SQH'. They have comparable security bounds: $k/|R|$, $|R| = p$ for SQH' and $2k/|R|$, $|R| = 2^n$ for CBH'.

5 Construction Based on SPN Network

In this section, we describe how to construct secure universal hash function by applying substitution permutation network (SPN) constructions to Theorem 2. This allows us to construct a provably secure MAC from block cipher components (as in Section 2) but whose speed is several times faster than CBC-MAC. First we will define the components needed in our hash function.

Let $s : GF(2)^t \rightarrow GF(2)^t$ be a bijective S-box. Define $S : [GF(2)^t]^r \rightarrow [GF(2)^t]^r$ as

$$S(z_1, \dots, z_r) = (s(z_1), \dots, s(z_r))$$

where each $z_i \in GF(2)^t$ is a t -bit binary vector. Let $MDS : [GF(2)^t]^r \rightarrow [GF(2)^t]^r$ denote a maximal distance separable transform. I.e., if

$$MDS(z_1, \dots, z_r) = (y_1, \dots, y_r)$$

where each $z_i, y_i \in GF(2)^t$, then $wt(z_1, \dots, z_r) + wt(y_1, \dots, y_r) \geq r + 1$ when $(z_1, \dots, z_r) \neq (0, \dots, 0)$.

Definition 5. *We define the SPN-hash $h_x : [GF(2)^{tr}]^k \rightarrow GF(2)^{tr}$ by the function $h_x(m) \triangleq \bigoplus_{i=1}^k S(MDS(S(m_i + x_i)))$.*

Theorem 6. *In definition 5, let the S-box $s : GF(2)^t \rightarrow GF(2)^t$ have maximal differential Δ_s . Then the SPN-hash is an ϵ - $A\Delta$ universal family of hash functions where the upper bound ϵ can be approximated by $(\frac{\Delta_s}{2^t})^{r+1}$.*

Proof. By Theorem 2, ϵ is at most the differential probability of $S \circ MDS \circ S : GF(2)^{tr} \rightarrow GF(2)^{tr}$.

Suppose we have an input difference $a \neq 0$ and output difference b for $S \circ MDS \circ S$. At least one of the r S-boxes in the first layer will have a non-zero input difference. Since the S-boxes are bijective, the output difference of the first layer of S-boxes, which is the input difference to the MDS transform, is non-zero. By the spreading effect of the MDS, there will be a total of $r + 1$ S-boxes with

non-zero input difference in the two layers of S-boxes in $S \circ MDS \circ S$. This means that differential approximation has probability at most $\approx (\Delta_s/2^t)^{r+1}$. I.e.

$$\Pr_z[S(MDS(S(z))) \oplus S(MDS(S(z \oplus a))) = b] \leq \epsilon \approx (\Delta_s/2^t)^{r+1}.$$

By Theorem 2, this implies for any two message blocks $m \neq n$,

$$\Pr[h_x(m) \oplus h_x(n) = \delta] \leq \epsilon \approx (\Delta_s/2^t)^{r+1}.$$

Note that we have used the term ‘‘approximation’’ of the upper bound ϵ . This is because there may be more than one differential path to achieve a specified input-output difference because our function is similar to a 2-round block cipher, so we are bounding the differential characteristic probability. However, the discrepancy between the differential probability and differential characteristic probability is not so big in our case because there are only two rounds. In a block cipher with 16 or more rounds, there are a lot more differential paths which lead to the same input-output difference.

Block ciphers designers, when they need to ensure protection against differential cryptanalysis in practice, just compute the best differential characteristic probability of a differential path. If they can get the probability to be less than $2^{-\text{block size}}$, then they say the cipher is secure against differential cryptanalysis. In our case, we adopt a similar strategy where we say the SPN-hash is secure if the security bound ϵ is approximated by a number close to $2^{-\text{block size}}$. \square

5.1 Example: An Application to Construct AES-Like MAC

Let $t = 8, r = 16$ in Theorem 6 and let $s(x) = x^{-1}$ over $GF(2^8)$. Then $\Delta_s = 4$ and the SPN-hash is an ϵ -A Δ universal family of hash functions where ϵ can be approximated by $(4/256)^{17} = 2^{-102}$. This bound is relatively close to the optimal bound 2^{-128} .

The components used are similar to that used in AES, except we are using a bigger MDS. The complexity is comparable to 2 rounds of AES encryption since we are using 32 inverse S-boxes on $GF(2^8)$ and a linear transform in both cases. This is 5 times faster than AES-128 (10 rounds) and 7 times faster than AES-256 (14 rounds).

When we construct a MAC by computing the SPN-hash with $t = 8, r = 16$ on the message block to get a universal hash value followed by an AES encryption. This is 5 to 7 times faster than CBC-MAC based on AES. Moreover, our SPN-hash is parallelizable since each message block can be processed independently while CBC-MAC has to wait for a message block to finish processing before going to the next. Suppose we perform 4 block computations simultaneously, that will increase the speed to more than 20 times than that of AES-CBC-MAC.

References

1. D. Ash, I. Blake, and S. Vanstone, ‘‘Low complexity normal bases,’’ *Discrete Applied Mathematics*, vol. 25, pp. 191–210, 1989.
2. E. Biham and A. Shamir, ‘‘Differential cryptanalysis of DES-like cryptosystems,’’ *Journal of Cryptology*, vol. 4, no. 1, pp. 3–72, 1991.

3. J. Black, S. Halevi, H. Krawczyk, T. Krovetz and P. Rogaway, "UMAC: Fast and secure message authentication," *Advances in Cryptology — CRYPTO '99, LNCS 1666*, pp. 216–233, Springer-Verlag, 1999.
4. G. Brassard. "On computationally secure authentication tags requiring short secret shared keys," *Advances in Cryptology — CRYPTO '83*, pp. 79–86, Springer-Verlag, 1983.
5. L. Budaghyan, C. Carlet and A. Pott, "New Classes of Almost Bent and Almost Perfect Nonlinear Functions," *Proceedings of Workshop on Coding and Cryptography 2005*, pp. 306–315, 2005.
6. J. L. Carter and M. N. Wegman, "Universal classes of hash functions," *Journal of Computer and System Sciences*, vol. 18, no. 2, pp. 143–154, 1979.
7. M. Etzel, S. Patel and Z. Ramzan, "Square hash: fast message authentication via optimized universal hash functions," *Advances in Cryptology — CRYPTO '99, LNCS 1666*, pp. 234–251, Springer-Verlag, 1999.
8. S. Halevi and H. Krawczyk, "MMH: Message authentication in software in the gbit/second rates," *Fast Software Encryption, FSE '97, LNCS 1267*, pp. 172–189, Springer-Verlag, 1997.
9. T. Helleseeth, "Correlation of m-sequences and related topics," In *Sequences and their Applications SETA '98*, pp. 49–66, 1999.
10. T. Helleseeth and T. Johansson, "Universal hash functions from exponential sums over finite fields and Galois rings," *Advances in Cryptology — CRYPTO '96, LNCS 1109*, pp. 31–44, Springer-Verlag, 1996.
11. S.-H. Heng and K. Kurosawa, "Square hash with a small key size," *Eighth Australasian Conference on Information Security and Privacy — ACISP '03, LNCS 2727*, pp. 522–531, Springer-Verlag, 2003.
12. K. Khoo and S.-H. Heng, "Universal Hash Functions over $GF(2^n)$," *Proceedings of 2004 IEEE International Symposium on Information Theory — ISIT 2004*, pp. 205, IEEE Press, 2004.
13. H. Krawczyk, "LFSR-based hashing and authentication," *Advances in Cryptology — CRYPTO '94, LNCS 839*, pp. 129–139, Springer-Verlag, 1994.
14. H. Krawczyk, "New hash functions for message authentication," *Advances in Cryptology — EUROCRYPT '95, LNCS 921*, pp. 301–310, Springer-Verlag, 1995.
15. V. Shoup, "On fast and provably secure message authentication based on universal hashing," *Advances in Cryptology — CRYPTO '96, LNCS 1109*, pp. 313–328, Springer-Verlag, 1996.
16. D. R. Stinson, "On the connections between universal hashing, combinatorial designs and error-correcting codes," *Congressus Numerantium 114*, pp. 7–27, 1996.
17. M. N. Wegman and J. L. Carter, "New hash functions and their use in authentication and set equality," *Journal of Computer and System Sciences*, vol. 22, no. 3, pp. 265–279, 1981.

Analysis of Fast Blockcipher-Based Hash Functions

Martin Stanek*

Department of Computer Science,
Faculty of Mathematics, Physics and Informatics,
Comenius University, Mlynská dolina, 842 48 Bratislava, Slovak Republic
`stanek@dcs.fmph.uniba.sk`

Abstract. An important property of a hash function is the performance. We study fast iterated hash functions based on block ciphers. These hash functions and their compression functions are analyzed in the standard black-box model. We show an upper bound on rate of any collision resistant hash function. In addition, we improve known bound on the rate of collision resistant compression functions.

Keywords: Hash functions, provable security, black-box model.

1 Introduction

Cryptographic hash functions are basic building blocks in many security constructions – digital signatures, message authentication codes, etc. Almost all modern hash functions are built by iterating a compression function according to the Merkle-Damgård paradigm [3, 6]. Moreover, these compression functions can be based on some underlying block cipher (we will use word ‘blockcipher’ in the paper).

The first systematic study of 64 blockcipher-based hash functions was done by Preneel, Govaerts, and Vandewalle [8]. Subsequently, Black, Rogaway, and Shrimpton [2] analyzed these constructions in the black-box model and proved that 20 of them are collision resistant up to the birthday-attack bound.

An important property of a hash function is the performance. Therefore, one would like to maximize the rate of a hash function – the number of message blocks processed with one blockcipher transformation. All classical blockcipher-based constructions [2, 8] process one message block with one transformation, thus they are rate-1. The “high-rate” compression functions, where single blockcipher transformation processes several message blocks, were studied in [7]. We showed that natural rate-2 generalizations of compression functions from [8] are not collision resistant.

Another way to design fast hash functions is to use keys from a small fixed set of keys in all block cipher transformations, thus enabling a pre-scheduling of keys in advance. Classical collision resistant constructions require rekeying for every message block. Recently, Black, Cochran, and Shrimpton [1] showed, that

* Supported by VEGA 1/3106/06.

it is impossible to construct (blockcipher-based) provably secure rate-1 iterated hash function that use small fixed set of keys.

Our Contribution. We analyze high-rate (blockcipher-based) iterated hash functions, and their underlying compression functions in the black-box model. Our contribution is twofold:

1. We prove, in the black-box model, an upper bound on the rate of collision resistant compression functions. The bound is an improvement of the result obtained in [7]. As a corollary we show that the rate must be 1 for certain classes of compression functions, in order to achieve their collision resistance.
2. We prove an upper bound on the rate of collision resistant iterated hash functions. Since the collision resistance of compression function is a sufficient [3, 6] but not necessary [2] condition for the collision resistance of a hash function, the upper bound obtained for compression functions is not directly applicable to hash functions.

The properties of some families of hash function constructions were analyzed in [4, 5], focussing on double length hash functions. Particularly, Knudsen and Lai presented attacks on all double block length hash functions of rate-1 [5]. However, their analysis covers double block length hash functions of special form. We analyze more general (and slightly different) model of hash/compression functions. Our model covers all hash functions constructed from compression functions that take r blocks of a message and process them using exactly one encryption transformation.

The paper is structured as follows. Section 2 contains notions and definitions used in the paper. The analysis of high-rate compression functions is presented in Section 3. We show the upper bound on the rate of iterated hash functions in Section 4.

2 Background and Definitions

The notation used in the paper follows closely the notation introduced in [1, 2]. Let $V_m = \{0, 1\}^m$ be a set of all m -ary binary vectors, and let V_m^* be a set of all binary strings obtained as a concatenation of (zero or more) elements from V_m . Let k and n be positive integers. A blockcipher is a function $E : V_k \times V_n \rightarrow V_n$, where for each key $K \in V_k$, the function $E_K(\cdot) = E(K, \cdot)$ is a permutation on V_n . Let $\text{Bloc}(k, n)$ be the set of all blockciphers $E : V_k \times V_n \rightarrow V_n$. The inverse of the blockcipher E is denoted by E^{-1} .

A (blockcipher-based) compression function is a function $f : \text{Bloc}(k, n) \times (V_a \times V_b) \rightarrow V_c$, where a , b , and c are positive integers such that $a + b \geq c$. We write the first argument as a superscript of compression function, i.e. $f^E(\cdot, \cdot) = f(E, \cdot, \cdot)$. An iterated hash of compression function $f : \text{Bloc}(k, n) \times (V_a \times V_b) \rightarrow V_a$ is the hash function $H : \text{Bloc}(k, n) \times V_b^* \rightarrow V_a$ defined by $H^E(m_1 \dots m_l) = h_l$, where $h_i = f^E(h_{i-1}, m_i)$ and h_0 is a fixed element from V_a . We set $H^E(\varepsilon) = h_0$ for an empty string ε . We often omit superscripts E to f and H , when the

blockcipher is known from the context. If the computation of $f^E(h, m)$ uses e evaluations of E then f (and its iterated hash H) is rate- r , where $r = (b/n)/e$. Often $n \mid b$, and the rate represents the average number of message blocks processed by single E transformation. For example, for $b/n = 2$ and $e = 1$ we get rate-2 compression/hash function.

Black-Box Model. An adversary A is given access to oracles E and E^{-1} where E is a blockcipher. We write these oracles as superscripts, i.e. $A^{E, E^{-1}}$. We omit the superscripts when the oracles are clear from the context. The adversary’s task is attacking the collision resistance of a hash function H . We measure the adversary’s effort of finding a collision as a function of the number of E or E^{-1} queries the adversary makes. Notice that we assume an information-theoretic adversary, i.e. the computational power of the adversary is not limited in any way.

Attacks in this model treat the blockcipher as a black box. The only structural property of the blockcipher captured by the model is the invertibility. The model cannot guarantee the security of blockcipher-based hash functions instantiated with blockciphers having significant structural properties (e.g. weak keys). On the other hand, the black-box model is stronger than treating the blockcipher as a random function, because of the adversary’s ability to compute E^{-1} .

We define the advantage of an adversary in finding collisions in a compression function $f : \text{Bloc}(k, n) \times (V_a \times V_b) \rightarrow V_c$. Naturally (h, m) and (h', m') collide under f if they are distinct and $f^E(h, m) = f^E(h', m')$. We also take into account a collision with empty string, i.e. producing (h, m) such that $f^E(h, m) = h_0$. We look at the number of queries that the adversary makes, and compare this with the probability of finding a collision.

The experiment of choosing a random element x from the finite set S will be denoted by $x \stackrel{\$}{\leftarrow} S$.

Definition 1 (Collision resistance of a compression function [2]). *Let f be a blockcipher-based compression function, $f : \text{Bloc}(k, n) \times (V_a \times V_b) \rightarrow V_c$. Fix a constant $h_0 \in V_c$ and an adversary A . Then the advantage of finding collisions in f is the probability*

$$\text{Adv}_f^{\text{comp}}(A) = \Pr \left[E \stackrel{\$}{\leftarrow} \text{Bloc}(k, n); ((h, m), (h', m')) \leftarrow A^{E, E^{-1}} : \right. \\ \left. (h, m) \neq (h', m') \wedge f^E(h, m) = f^E(h', m') \vee f^E(h, m) = h_0 \right]$$

For $q \geq 0$ we write $\text{Adv}_f^{\text{comp}}(q) = \max_A \{ \text{Adv}_f^{\text{comp}}(A) \}$ where the maximum is taken over all adversaries that ask at most q oracle (E or E^{-1}) queries.

Definition 2 (Collision resistance of a hash function [2]). *Let H be a blockcipher-based hash function, and let A be an adversary. Then the advantage of finding collisions in H is the probability*

$$\text{Adv}_H^{\text{coll}}(A) = \Pr \left[E \stackrel{\$}{\leftarrow} \text{Bloc}(k, n); (M, M') \leftarrow A^{E, E^{-1}} : \right. \\ \left. M \neq M' \wedge H^E(M) = H^E(M') \right]$$

For $q \geq 0$ we write $\mathbf{Adv}_H^{\text{coll}}(q) = \max_A \{\mathbf{Adv}_H^{\text{coll}}(A)\}$ where the maximum is taken over all adversaries that ask at most q oracle (E or E^{-1}) queries.

The following theorem forms a basis for a construction of iterated hash functions (Merkle-Damgård paradigm). It shows that the collision resistance of compression function is sufficient for the collision resistance of its iterated hash function.

Theorem 1 (Merkle-Damgård [3, 6]). *Let $f : \text{Bloc}(k, n) \times V_n \times V_n \rightarrow V_n$ be a compression function and let H be an iterated hash of f . Then $\mathbf{Adv}_H^{\text{coll}}(q) \leq \mathbf{Adv}_f^{\text{comp}}(q)$ for any $q \geq 1$.*

The birthday attack is a generic collision-finding attack on a compression/hash function. The advantage of the birthday attack is $\Theta(q^2/2^n)$, where q is the number of evaluations of the compression/hash function and n is the length of the output. Usually, a compression function f (hash function H) is called collision resistant up to the birthday-attack bound, or simply collision resistant if $\mathbf{Adv}_f^{\text{comp}}(q) = O(q^2/2^n)$ ($\mathbf{Adv}_H^{\text{coll}}(q) = O(q^2/2^n)$). Since the birthday attack is always applicable, these equations can be stated equivalently as $\mathbf{Adv}_f^{\text{comp}}(q) = \Theta(q^2/2^n)$ ($\mathbf{Adv}_H^{\text{coll}}(q) = \Theta(q^2/2^n)$).

Model of High-Rate Compression/Hash Functions. We use the model of high-rate compression/hash functions proposed in [7]. Let f be a compression function. The model assumes the following:

- The computation of $f(h, m)$ uses a single evaluation of E .
- The length of m is an integer multiple of the E 's block length n , i.e. $b = rn$ where $r \geq 1$. Thus, the rate of f is r .

The computation of the compression function $f : \text{Bloc}(k, n) \times (V_a \times V_{rn}) \rightarrow V_a$ can be defined as:

$$f(h, m) = f_3(h, m, E_{f_2(h,m)}(f_1(h, m))), \tag{1}$$

where $f_1 : V_a \times V_{rn} \rightarrow V_n$, $f_2 : V_a \times V_{rn} \rightarrow V_k$, and $f_3 : V_a \times V_{rn} \times V_n \rightarrow V_a$ are arbitrary functions. When convenient we express m as a concatenation of n -bit blocks. These r blocks are denoted by $m^{(1)}, \dots, m^{(r)}$. A high-rate hash function is an iterated hash of a compression function f (see Fig. 1).

The model covers all compression functions that take r blocks of a message and process them using exactly one encryption transformation E . Notice that all rate-1 schemes from [8] are special instances of the model.

3 Analysis of High-Rate Compression Functions

The rate of any collision resistant compression function in the model is upper bounded by $1 + k/n$ [7]. The following theorem considers the number of oracle queries that the adversary makes, and improves this upper bound.

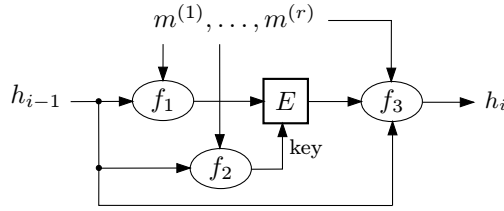


Fig. 1. Model of high-rate compression/hash function

Theorem 2. Let $f : E \in \text{Bloc}(k, n) \times (V_a \times V_{rn}) \rightarrow V_a$ be a compression function defined by (1). Let $q \geq 1$ denotes the number of oracle queries (E or E^{-1}). Let $r > 1 + \frac{k - \lg q}{n}$. Then $\text{Adv}_f^{\text{comp}}(q) = 1$.

Proof. Let $\mathcal{D}_f = V_n \times V_{rn}$ be the domain of the compression function f . We partition this set into distinct sets $D_{X,K}$ (for $X \in V_n, K \in V_k$) defined as follows:

$$D_{X,K} = \{(h, m) \mid f_1(h, m) = X, f_2(h, m) = K\}.$$

We describe an adversary A that asks exactly q oracle queries. The adversary tries to find collisions in the following way:

1. A finds q largest sets $D_{X,K}$ – we denote them $D_1 = D_{X_1, K_1}, \dots, D_q = D_{X_q, K_q}$. Notice that no oracle queries are required for computing $|D_{X,K}|$, for arbitrary X and K .
2. A asks q queries, and computes $Y_i = E_{K_i}(X_i)$, for $i = 1, \dots, q$.
3. A finds the collision in the set $D = \bigcup_{i=1}^q D_i$, i.e. $(h, m), (h', m') \in D$ such that $(h, m) \neq (h', m')$ and $f(h, m) = f(h', m')$. Since any $(h, m) \in D$ is a member of some D_i , and the value $Y_i = E_{K_i}(X_i)$ is already known, the computation of $f(h, m)$ requires no further oracle queries.

Steps 1 and 3 have an exponential complexity – recall that adversaries in the black-box model are computationally unlimited. What counts is the number of oracle queries.

It remains to show that A finds the collision in step 3 with probability 1. Since D_1, \dots, D_q are the largest among the sets $D_{X,K}$, we can estimate the lower bound of $|D|$ according to the pigeonhole principle:

$$|D| \geq \frac{|\mathcal{D}_f|}{|V_n \times V_k|/q} = \frac{q \cdot 2^{a+rn}}{2^{n+k}} = q \cdot 2^{a+(r-1)n-k}.$$

The range of the compression function f is V_a . Hence, if $|D| > 2^a$ the adversary succeeds in finding collision with probability 1. We get

$$q \cdot 2^{a+n(r-1)-k} > 2^a \iff r > 1 + \frac{k - \lg q}{n}.$$

Adversary A finds a collision in the compression function f with probability 1 asking exactly q oracle (E) queries during the attack. Thus, $\text{Adv}_f^{\text{comp}}(q) = 1$. \square

The constructions of compression functions from the blockcipher often assume equal key and block lengths, i.e. $k = n$. Similarly, the output of a compression function has usually the same length as the block, i.e. $a = n$. Let us apply Theorem 2 with $q = n$ queries in this scenario. We get $\mathbf{Adv}_f^{\text{comp}}(n) = 1$ for $r > 1 + \frac{n - \lg n}{n} = 2 - \frac{\lg n}{n}$. Since r is an integer, this is equivalent to $r > 1$. Comparing the advantage of the adversary from Theorem 2 with the advantage of the birthday attack ($\Theta(n^2/2^n)$, for $q = n$) yields the following result:

Corollary 1. *Let $f : \text{Bloc}(n, n) \times (V_n \times V_{rn}) \rightarrow V_n$ be a compression function defined by (1). Let $r > 1$. Then f is not collision resistant.*

Remark 1. The result of non-existence of collision resistant rate-2 generalizations of classical compression functions [7] now easily follows from Corollary 1.

Theorem 2 shows an upper bound depending on the number of oracle queries. Following corollary relates the upper bound to the lengths n , k , and a .

Corollary 2. *Let $f : \text{Bloc}(k, n) \times (V_a \times V_{rn}) \rightarrow V_a$ be a compression function defined by (1). Let $0 \leq \varepsilon < 1$ be an arbitrary constant. Let $r > 1 + \frac{k - \varepsilon a/2}{n}$. Then f is not collision resistant.*

Proof. Let $q = 2^{\varepsilon a/2}$. The advantage of the birthday attack with q evaluations of f is $\Theta(q^2/2^a) = \Theta(2^{a(\varepsilon-1)})$. Recall that exactly one oracle query (computation of E) is needed for evaluation of f in our model of high-rate compression functions. The assumption $r > 1 + (k - \lg q)/n$ of Theorem 2 is equivalent to $r > 1 + \frac{k - \varepsilon a/2}{n}$ for our q . The theorem yields $\mathbf{Adv}_f^{\text{comp}}(q) = 1$. This advantage is asymptotically greater than the advantage of the birthday attack. Thus, the compression function f is not collision resistant. \square

Remark 2. We considered compression functions in the model specified in Section 2. Our results do not rule out the possibility that collision resistant high-rate compression functions exist outside the model.

It is well known that the collision resistance of a compression function is a sufficient but not necessary condition for the collision resistance of its iterated hash. Thus, the upper bounds obtained in Theorem 2 and Corollary 2 are not directly applicable to hash functions. We study the possibility of high-rate collision resistant hash functions in the following section.

4 Upper Bound on the Rate of Hash Functions

We consider hash functions in the model of high-rate compression/hash functions specified in Section 2. Recall that the iterated hash H based on a compression function f is computed as follows:

$$\begin{aligned}
 H(m_1 \dots m_l) = h_l & : & H(\varepsilon) = h_0 \\
 & & h_i = f(h_{i-1}, m_i), \quad i \geq 1,
 \end{aligned}$$

where h_0 is a fixed value, and ε denotes an empty string.

Theorem 3. Let $f : \text{Bloc}(k, n) \times (V_a \times V_{rn}) \rightarrow V_a$ be a compression function defined by (1). Let $H : V_{rn}^* \rightarrow V_a$ be an iterated hash based on f . Let $q \geq 1$ denotes the number of oracle queries. Let $r > 1 + \frac{k+a/q}{n}$. Then $\text{Adv}_H^{\text{coll}}(q) = 1$.

Proof. We describe an adversary A that asks at most q oracle queries, and finds a collision in H . The adversary builds a directed tree T with labeled nodes and edges (see Fig. 2):

- edges are labeled by message blocks (from V_{rn}),
- nodes are labeled by “intermediary” hash values (from V_a),
- the root is labeled by h_0 ,
- for every edge $h \xrightarrow{m} h'$ (h, h', m are labels) following relation holds: $f(h, m) = h'$.

The tree T describes computations of hash function H on various messages. Every path starting in the root ends in the node labeled by value $H(m)$, where m is message obtained by concatenating blocks (edges labels) along the path. Multiple nodes labeled by the same value are allowed in T .

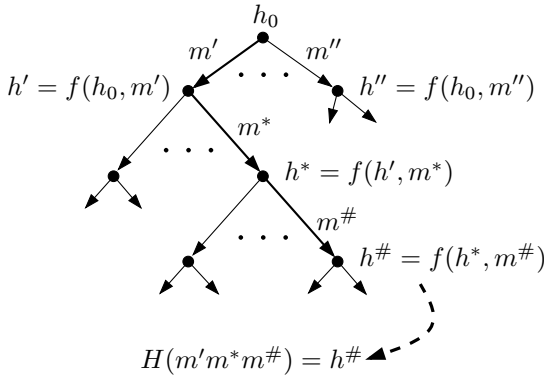


Fig. 2. Labeling nodes and edges in T

Adversary A starts with single root node labeled by h_0 . Let

$$D_{X,K}^{(0)} = \{(h_0, m) \mid m \in V_{rn}, f_1(h_0, m) = X, f_2(h_0, m) = K\},$$

for any $X \in V_n$ and $K \in V_k$. Adversary enumerates $|D_{X,K}^{(0)}|$ for all 2^{n+k} values X and K , and finds the largest (maximal) one. Notice that no oracle query is needed for this computation. Let $D_{\max}^{(0)}$ be the largest set. Its size can be bounded according to the pigeonhole principle:

$$|D_{\max}^{(0)}| \geq \frac{|\bigcup_{X,K} D_{X,K}^{(0)}|}{2^{n+k}} = \frac{2^{rn}}{2^{n+k}} = 2^{n(r-1)-k}.$$

Since for every $(h_0, m) \in D_{\max}^{(0)}$ the values $f_1(h_0, m)$ and $f_2(h_0, m)$ are fixed, A can compute $f(h_0, m)$ for all pairs from $D_{\max}^{(0)}$ with single oracle query. For every $(h_0, m) \in D_{\max}^{(0)}$, A inserts a new node labeled by $f(h_0, m)$ and edge $h_0 \xrightarrow{m} f(h_0, m)$ into T . The adversary distinguishes two cases:

1. There are equally labeled nodes in depth 1. Thus, two different messages (edge labels) have the same hash – A outputs this collision and stops.
2. All nodes in depth 1 have distinct labels – there are at least $2^{n(r-1)-k}$ distinct hash values. In this case, A continues the attack (and extends T to depth 2, 3, etc.).

The adversary runs the following cycle for $i = 1, \dots, q - 1$:

0. (*invariant*) All nodes in depth i have distinct labels.
1. Let

$$D_{X,K}^{(i)} = \{(h, m) \mid h \in D_{\max}^{(i-1)}, m \in V_{rn}, f_1(h, m) = X, f_2(h, m) = K\},$$

for any $X \in V_n$ and $K \in V_k$. Adversary A enumerates $|D_{X,K}^{(i)}|$ for all 2^{n+k} values X and K (again, no oracle queries needed), and finds the largest one – $D_{\max}^{(i)}$. Its size can be bounded:

$$|D_{\max}^{(i)}| \geq \frac{\left| \bigcup_{X,K} D_{X,K}^{(i)} \right|}{2^{n+k}} = \frac{|D_{\max}^{(i-1)}|}{2^{n+k}} = \frac{2^{rn} \cdot 2^{i(n(r-1)-k)}}{2^{n+k}} = 2^{(i+1)(n(r-1)-k)}.$$

2. The adversary A computes $f(h, m)$ for all pairs from $D_{\max}^{(i)}$. Since for every $(h, m) \in D_{\max}^{(i)}$ the values $f_1(h, m)$ and $f_2(h, m)$ are fixed, A performs whole computation using single oracle query. For every $(h, m) \in D_{\max}^{(i)}$, A inserts a new node labeled by $f(h, m)$ and edge $h \xrightarrow{m} f(h, m)$ in T . Each added edge starts in depth i (thus, starting in the deepest leaves).
3. If there exist two equally labeled nodes in depth $i + 1$, the adversary found a collision. The messages along the paths from the root to these nodes must differ: If the nodes share common parent, the last blocks are different. In case of distinct parents we have distinct intermediate hash values (see the invariant) – so the messages up to these parents are different. Hence, A outputs the collision and stops.
4. If all nodes in depth $i+1$ have distinct labels, there are at least $2^{(i+1)(n(r-1)-k)}$ nodes with distinct labels. The adversary A continues the cycle.

Let us explore the situation in the last run of the cycle ($i = q - 1$). We know that $|D_{\max}^{(q-1)}| \geq 2^{q(n(r-1)-k)}$. There are only 2^a distinct hashes. Therefore, there must be equally labeled nodes in depth q if $2^{q(n(r-1)-k)} > 2^a$. This is equivalent to $r > 1 + \frac{k+a/q}{n}$.

Hence, adversary A succeeds in finding the collision in H with probability 1. Overall, A asks at most q oracle queries. Thus, $\mathbf{Adv}_H^{\text{coll}}(q) = 1$. □

Remark 3. The adversary A finds colliding messages of equal length. Therefore employing the message length in input encoding, such as MD-strengthening, does not affect the bound in Theorem 3.

Let us analyze (again) the situation for $k = a = n$. We apply Theorem 3 with $q = n$ oracle queries. We get $\text{Adv}_H^{\text{coll}}(n) = 1$ for $r > 1 + \frac{n+n/n}{n} = 2 + \frac{1}{n}$. Since r is an integer, this is equivalent to $r > 2$ (for $n > 1$). Comparing this advantage of the adversary with the advantage of the birthday attack ($\Theta(n^2/2^n)$, for $q = n$) yields the following result:

Corollary 3. *Let $f : \text{Bloc}(n, n) \times (V_n \times V_{rn} \rightarrow V_n)$ be a compression function defined by (1). Let $H : V_{rn}^* \rightarrow V_n$ be an iterated hash based on f . Let $r > 2$. Then H is not collision resistant.*

The following corollary relates the upper bound on r to the lengths n , k , and a .

Corollary 4. *Let $f : \text{Bloc}(k, n) \times (V_a \times V_{rn}) \rightarrow V_a$ be a compression function defined by (1). Let $H : V_{rn}^* \rightarrow V_a$ be an iterated hash based on f . Let $0 \leq \varepsilon < 1$ be an arbitrary constant. Let $r > 1 + \frac{k+a/(2^{\varepsilon a/2})}{n}$. Then H is not collision resistant.*

Proof. The proof is similar to the proof of Corollary 2. We apply Theorem 3 for $q = 2^{\varepsilon a/2}$ oracle queries and compare the advantage of the adversary (i.e. 1) with the advantage of the birthday attack (i.e. $\Theta(2^{a(\varepsilon-1)})$). \square

5 Conclusion

We analyzed high-rate compression/hash functions in the black-box model. Table 1 summarizes main results.

Table 1. Upper bounds on the rate of collision resistant compression/hash functions

	compression function	hash function
general case	$1 + \frac{k - \varepsilon a/2}{n}$	$1 + \frac{k+a/(2^{\varepsilon a/2})}{n}$
$a = k = n$	1	2

These upper bounds narrow the space for a potential high-rate collision resistant hash function – at least they show where not to search for such functions. Probably the most interesting open problem is whether any collision resistant hash function of rate greater than 1 actually exists.

References

1. Black, J., Cochran, M., Shrimpton, T.: On the Impossibility of Highly-Efficient Blockcipher-Based Hash Functions, In *Advances in Cryptology – Eurocrypt ’05*, LNCS 3494, Springer-Verlag, 2005.

2. Black, J., Rogaway, P., Shrimpton, T.: Black-box analysis of the block-cipher-based hash-function constructions from PGV, In *Advances in Cryptology – CRYPTO '02*, LNCS 2442, Springer-Verlag, 2002.
3. Damgård, I.: A design principle for hash functions, In *Advances in Cryptology – CRYPTO '89*, LNCS 435, Springer-Verlag, 1990.
4. Hohl, W., Lai, X., Meier, T., Waldvogel, C.: Security of iterated hash function based on block ciphers, In *Advances in Cryptology – CRYPTO '93*, LNCS 773, Springer-Verlag, 1994.
5. Knudsen, L., Lai, X.: New attacks on all double block length hash functions of hash rate 1, including parallel-DM, In *Advances in Cryptology – Eurocrypt '94*, LNCS 950, Springer-Verlag, 1994.
6. Merkle, R.: One way hash functions and DES, In *Advances in Cryptology – CRYPTO '89*, LNCS 435, Springer-Verlag, 1990.
7. Ostertág, R., Stanek, M.: On High-Rate Cryptographic Compression Functions, *Cryptology ePrint Archive*, Report 2005/152, <http://eprint.iacr.org/>, 2005.
8. Preneel, B., Govaerts, R., Vandewalle, J.: Hash functions based on block ciphers: A synthetic approach, In *Advances in Cryptology – CRYPTO '93*, LNCS 773, Springer-Verlag, 1994.

Application of LFSRs for Parallel Sequence Generation in Cryptologic Algorithms

Sourav Mukhopadhyay and Palash Sarkar

Cryptology Research Group, Applied Statistics Unit,
Indian Statistical Institute, 203, B.T. Road, Kolkata, India 700108
{sourav_t,palash}@isical.ac.in

Abstract. We consider the problem of efficiently generating sequences in hardware for use in certain cryptographic algorithms. The conventional method of doing this is to use a counter. We show that sequences generated by linear feedback shift registers (LFSRs) can be tailored to suit the appropriate algorithms. For hardware implementation, this reduces both time and chip area. As a result, we are able to suggest improvements to the design of DES Cracker built by the Electronic Frontier Foundation in 1998; provide an efficient strategy for generating start points in time-memory trade/off attacks; and present an improved parallel hardware implementation of a variant of the counter mode of operation of a block cipher.

Keywords: DES Cracker, TMTO, Counter Mode of Operation, LFSR.

1 Introduction

Consider the following cryptologic algorithms which require the generation of a sequence of s -bit vectors.

Exhaustive Search: In this case, the search space consists of all elements of $\{0, 1\}^s$ and the algorithm must consider each element of this set. Exhaustive search algorithms like the DES Cracker [1] employ a high degree of parallelism. Hence, the requirement is to generate in parallel a set of pairwise disjoint sequences of s -bit vectors whose union is the set $\{0, 1\}^s$.

Time-Memory Trade-Off (TMTO): This is also a generic search method. The pre-computation phase of such algorithms require the generation of parallel independent (pseudo)-random sequences of s -bit values.

Counter Mode of Operation: This is a mode of operation of a block cipher, which converts the block cipher into an additive stream cipher. In this mode of operation, one requires to generate a long non-repeating sequence of s -bit values.

The first two are cryptanalytic algorithms, while the third one is a cryptographic algorithm. Implementations of the above three algorithms use a counter to generate the required sequences. While this is intuitively simple, it is not the best possible option for hardware implementation.

In this paper, we explore the possibility of using sequences obtained from linear feedback shift registers (LFSRs) for the hardware implementation of the above algorithms. In each case, we show how LFSR sequences can be tailored for use in the respective algorithm. Replacing counters by LFSRs in hardware implementation has the following two advantages.

Time: The next state of an LFSR can be obtained in one clock. For a counter, we need to add one to the current value. Addition requires much more time. We also show that parallel generation of pairwise disjoint subsequences can be done very efficiently.

Chip Area: Implementing the next state function of an LFSR in hardware requires only a few XOR gates. In contrast, sophisticated carry-look-ahead adders require significantly more circuitry. Consequently, replacing adders by LFSRs will reduce the required chip area.

A combination of the above two effects can lead to a significant improvement in price-performance ratio. This leads us to suggest changes to the DES Cracker [1] which simplify the design as well as reduce the time; to provide an efficient strategy for generating start points in hardware implementation of TMTO algorithms; and finally, to present a new variant of the classical counter mode of operation of a block cipher. This new variant has a more efficient parallel hardware implementation. The first work to suggest the use of LFSRs in exhaustive key search was by Wiener [13]. (This fact was pointed out to us by David Wagner.) See [10] for more details.

2 LFSR Preliminaries

A binary linear feedback shift register (LFSR) of length s is an s -bit register. Let at time $t \geq 0$, the content of stage i be $a_i^t \in \{0, 1\}$ for $0 \leq i \leq s-1$. Then the state at time t is given by the vector $S_t = (a_{s-1}^{(t)}, a_{s-2}^{(t)}, \dots, a_0^{(t)})$. The state at time $t+1$ is given by the vector $S_{t+1} = (a_{s-1}^{(t+1)}, a_{s-2}^{(t+1)}, \dots, a_0^{(t+1)})$, where $a_i^{(t+1)} = a_{i+1}^{(t)}$ for $0 \leq i \leq s-2$; and $a_{s-1}^{(t+1)} = c_0 a_{s-1}^{(t)} c_1 a_{s-2}^{(t)} \oplus \dots \oplus c_{s-1} a_0^{(t)}$. The values c_0, \dots, c_{s-1} are constant bits and the polynomial $p(x) = x^s \oplus c_{s-1} x^{s-1} \oplus c_{s-2} x^{s-2} \oplus \dots \oplus c_1 x \oplus c_0$ over $GF(2)$ is called the connection polynomial of the LFSR. The behaviour of an LFSR is described by the sequence S_0, S_1, \dots of s -bit vectors and is completely determined by the state S_0 and the polynomial $p(x)$.

Maximal Length LFSR: It is well known that if $p(x)$ is a primitive polynomial, then for any non-zero s -bit vector S_0 , the sequence $S_0, S_1, S_2, \dots, S_{2^s-2}$ consists of all the $2^s - 1$ non-zero s -bit vectors. An LFSR which has this property is called a maximal length LFSR.

Matrix Representation: There is another way to view an LFSR sequence, which will be useful to us later. The next state S_{t+1} is obtained from state S_t by a linear transformation and hence we can write $S_{t+1} = S_t M$, where M is an $s \times s$ matrix whose characteristic polynomial is $p(x)$. Extending this, we can write $S_t = S_0 M^t$.

Thus, knowing M^t , we can directly jump from S_0 to S_t without going through the intermediate states. For any fixed value of $t < 2^s - 1$, computing the matrix exponentiation M^t can be done using the usual square and multiply method and requires at most $2 \log t \leq 2s$ matrix multiplications.

Implementation: Implementing an LFSR in hardware is particularly efficient. Such an implementation requires s flip-flops and $\text{wt}(p(x)) - 1$ many 2-input XOR gates, where $\text{wt}(p(x))$ is the number of non-zero coefficients in $p(x)$. With this hardware cost, the next s -bit state is obtained in one clock. For maximal length LFSR, one requires $p(x)$ to be primitive. It is usually possible to choose $p(x)$ to be of very low weight, either a trinomial or a pentanomial. Thus, an s -bit maximal length LFSR provides a fast and low cost hardware based method for generating the set of all non-zero s -bit vectors. Software generation of an LFSR sequence is in general not as efficient as in hardware. On a machine which supports w -bit words, the next s -bit state of an LFSR can be obtained using $(\text{wt}(p(x)) - 1)s/w$ XOR operations (see [3]).

2.1 LFSRs Versus Counters

The set of all s -bit vectors can be identified with the set of non-negative integers less than 2^s . In certain cryptologic algorithms, the requirement is to generate a sequence of non-negative integers with the only condition that no value should repeat. One simple way of doing this is to generate integers $0, 1, 2, \dots$ using a counter. While intuitively simple, this is not the only method of generating non-repeating sequence. One can use an s -bit maximal length LFSR to generate the sequence S_0, S_1, \dots , which is also non-repeating. We next discuss the relative advantages of LFSR and counter sequences with respect to hardware implementation.

Implementing a counter which can count from 0 upto $2^s - 1$ requires an s -bit register and an adder. The task required of the adder is to add one to the current state of the register. Due to carry propagation, the simplest adder implementation will require s clocks in the worst case (and $s/2$ clocks in the average case) to generate the next value. More sophisticated carry-look-ahead adders can reduce the number of clocks but the circuitry becomes significantly more complicated and costlier. In contrast, for LFSR sequences, apart from the s -bit register, we require only $\text{wt}(p(x)) - 1$ many 2-input XOR gates and the next s -bit state is obtained in one clock cycle.

Another advantage is that of scalability. The main cost of implementing an LFSR is the register and the interconnections. The number of XOR gates can usually be taken to be either two or four and can be assumed to be less than ten for all values of s . Thus, the cost of implementing an LFSR scales linearly with the value of s . On the other hand, the cost of implementing an adder circuit scales quadratically with the value of s .

Hence, using an LFSR in place of a counter leads to significantly lower hardware cost and also provides a faster method of generating a non-repeating sequence. Additionally, for certain applications, the requirement is to generate a

pseudorandom sequence of non-negative integers. In such cases, the only option is to use an LFSR sequence.

3 Parallel Sequence Generation

Consider the following problem.

- Generate n parallel and pairwise disjoint sequences of s -bit strings such that the union of these n sequences is the set of all (non-zero) s -bit strings.

We provide a simple LFSR based strategy for solving the above problem. Let $L = (s, p(x))$ be an s -bit LFSR where $p(x)$ is a primitive polynomial of degree s over $GF(2)$. Let $2^s - 1 = \tau \times n + r = (\tau + 1)r + \tau(n - r)$ where $0 \leq r < n$. Let $n_1 = n - r, n_2 = r$ and note that $\tau = \lfloor \frac{2^s - 1}{n} \rfloor$. Let S_0 be any nonzero s -bit string and for $t \geq 1$, we define $S_t = S_0 M^t$, where M is the state transition matrix of L . Further, let $T_0 = S_{n_1 \tau}$ and for $t \geq 1, T_t = T_0 M^t = T_{t-1} M$. Also let $\tau' = \lceil \frac{2^s - 1}{n} \rceil$. Define n sequences as follows.

$$\begin{array}{ll}
 \mathcal{S}_0 : & S_0, S_1, \dots, S_{\tau-1}; & \mathcal{T}_0 : & T_0, T_1, \dots, T_{\tau-1}; \\
 \mathcal{S}_1 : & S_{\tau}, S_{\tau+1}, \dots, S_{2\tau-1}; & \mathcal{T}_1 : & T_{\tau}, T_{\tau+1}, \dots, T_{2\tau-1}; \\
 \vdots & & \vdots & \\
 \mathcal{S}_{n_1-1} : & S_{(n_1-1)\tau}, \dots, S_{n_1\tau-1}; & \mathcal{T}_{n_2-1} : & T_{(n_2-1)\tau}, \dots, T_{n_2\tau-1}.
 \end{array} \tag{1}$$

The \mathcal{S} sequences are of length τ , while the \mathcal{T} sequences are of length $\tau' \geq \tau$. Note that, $T_{n_2\tau-1} = T_0 M^{n_2\tau-1} = S_0 M^{n_1\tau} M^{n_2\tau-1} = S_0 M^{n_1\tau+n_2\tau-1} = S_0 M^{2^s-2} = S_{2^s-2}$. Since $p(x)$ is primitive, the sequence $S_0, S_1, \dots, S_{n_1\tau-1}, T_0, T_1, \dots, T_{n_2\tau-1}$ consists of all non-zero s -bit vectors. This ensures that the sequences $\mathcal{S}_0, \mathcal{S}_1, \dots, \mathcal{S}_{n_1-1}, \mathcal{T}_0, \mathcal{T}_1, \dots, \mathcal{T}_{n_2-1}$ are pairwise disjoint. Thus, we obtain a solution to the problem mentioned above. We now consider the problem of actually generating the sequences in hardware.

Implementation: Let L_0, \dots, L_{n-1} be n implementations of the LFSR L . Hence, each L_i has $p(x)$ as its connection polynomial. The initial conditions for L_0, \dots, L_{n_1-1} are $S_0, S_{\tau}, \dots, S_{(n_1-1)\tau}$ respectively and the initial conditions for L_{n_1}, \dots, L_{n-1} are $T_0, T_{\tau}, \dots, T_{(n_2-1)\tau}$ respectively. At any point of time, the current states of the L_i 's provide the current values of the \mathcal{S} and the \mathcal{T} sequences. All the L_i 's operate in parallel, i.e., they are all clocked together and hence the next states of the \mathcal{S} and the \mathcal{T} sequences are generated in parallel. The total hardware cost for implementing the n LFSRs consists of $n \times s$ flip-flops and $n \times (\text{wt}(p(x) - 1))$ many 2-input XOR gates. With this minimal hardware cost, the parallel generation of the \mathcal{S} and the \mathcal{T} sequences become possible.

Obtaining the Initial Conditions: We explain how to obtain the initial condition for the n LFSRs. Let $M_1 = M^{\tau}$ and $M_2 = M^{\tau+1} = M \times M_1$. Then $S_{i\tau} = S_0 M^{i\tau} = S_{(i-1)\tau} \times M^{\tau} = S_{(i-1)\tau} \times M_1$. Now $T_0 = S_{(n_1-1)\tau} \times M_1$ and $T_{j\tau} = T_{(j-1)\tau} \times M_2$. Once we know M_1 and M_2 it is easy to find all the $S_{i\tau}$'s

and $T_{j\tau}$'s. Computing M_1 requires a matrix exponentiation which as mentioned before requires $2 \log \tau \leq 2s$ many matrix multiplications. Obtaining M_2 from M_1 requires one matrix multiplication. After M_1 and M_2 have been obtained, computing the initial conditions require a total of n many vector-matrix multiplications. These initial conditions are obtained once for all in an offline phase. These are then pre-loaded into the LFSRs and do not need to re-computed during the actual generation of the parallel sequences.

4 Application 1: The DES Cracker

In the design of DES cracker [1], a computer drives 2^{16} *search units*. The search units are parallel hardware units while the computer provides a central control software. The key space is divided into segments and each search unit searches through one segment. For each candidate key, a search unit does the following. Let k be the current candidate key. A search unit decrypts the first ciphertext using k and checks whether the resulting plaintext is "interesting". If yes, then it decrypts the second ciphertext using k and checks if it is also interesting. (The search unit considers a plaintext to be interesting if all its 8 bytes are ASCII.) If the both plaintexts are found to be interesting then the (key, plaintext) pair is passed to a computer to take the final decision. The search unit then adds one to k to obtain the next candidate key.

Recall that in DES, the message and cipher block size is 64 bits while the key size is 56 bits. In each search unit, a counter (and an adder) generates the candidate keys. A 32-bit counter is used to count through the bottom 32 bits of the key. The reason for using a 32-bit adder is that it is cheaper to implement than a 56-bit adder. The top 24 bits of the key are loaded onto the search unit by the computer. After completing 2^{32} keys with a fixed value of the 24 bits, a search unit sends a signal to the computer. The computer stops the chip; resets the key counter; puts a new value in the top 24 bits; and the search starts once more with this new 24-bit value.

4.1 LFSR Based Solution

We describe an alternative LFSR based solution for candidate key generation in the DES cracker. This solution is based on the parallel sequence generation described in Section 3. The number of parallel search units $n = 2^{16}$, while $s = 56$. Thus, $\tau = 2^{40} - 1$, $\tau' = 2^{40}$, $n_1 = 1$ and $n_2 = 2^{16} - 1$.

Choose the LFSR L such that $p(x)$ is the primitive pentanomial ($x^{56} + x^{22} + x^{21} + x + 1$). Choose S_0 to be an arbitrary non-zero 56-bit value and compute the values T_0, \dots, T_{n_2-1} using the method of Section 3. The total number of 56×56 binary matrix multiplications required is at most $2 \times s + 1 = 113$. Additionally, one has to compute a total of 2^{16} many multiplications of a 56-bit vector with a 56×56 binary matrix. Even with a straightforward software implementation, the entire computation can be completed within a few hours. The initial condition of the LFSR in the first search unit is set to S_0 , while the initial conditions for

the LFSRs in the other search units are set to $T_0, T_1, \dots, T_{n_2-1}$. Computing the initial conditions can be considered to be part of design stage activity.

In our design, each search unit of the DES cracker has its own implementation of L . This implementation requires n flip-flops and only four 2-input XOR gates. Each search unit now generates the candidate keys independently of the computer and also independently of each other. To obtain the next candidate key, it simply clocks its local LFSR once and uses the state of the LFSR as the candidate key. The first search unit does this for $\tau = 2^{40} - 1$ steps while the other search units do this for $\tau' = 2^{40}$ steps. This ensures that all non-zero keys are considered, with the all-zero key being considered separately.

4.2 Comparison to the Counter Based Solution

There are two ways in which the LFSR based solution improves over the counter based solution.

- There are 2^{16} search units. In the counter based solution, each search unit sends an interrupt signal to the computer after completing an assigned key segment. Thus, the computer needs to handle a total of 2^{24} interrupts from all the search units. In the LFSR based solution, candidate key generation is done solely by the search unit without any involvement from the computer.
- In the counter method, each search unit requires a 32-bit adder for a total of 2^{16} such adders. In contrast, in the LFSR based solution, the circuitry for generating the next candidate key consists of only 4 XOR gates per search unit. Thus, the adders of the counter based method take up significantly more chip area which could be utilised otherwise. One could either build more parallel search units at the same cost, or build the same number of search units at a lesser cost.

The above two factors can lead to a substantial improvement in the price-performance ratio of the DES cracker.

4.3 General Exhaustive Search

The LFSR based candidate key generation algorithm described above for DES cracker can easily be generalized to generate candidate keys for exhaustive search on any cryptographic algorithm. We need to choose the appropriate value of s (for example, for AES, $s = 128$) and a suitable primitive polynomial of degree s over $GF(2)$. Now given n , the number of parallel search units, we can apply the method of Section 3 to obtain the initial conditions of the local LFSR implementations of all the search units. This in effect divides the entire key space into disjoint subspaces, with each search unit searching through its allotted subspace.

5 Application 2: TMTO Pre-computation

TMTO [5, 11] can be considered to be a generic method for inverting a one-way function. This consists of two phases: pre-computation phase and online attack

phase. A set of table(s) is constructed during the pre-computation phase. The tables store keys in an off-line phase. In the online phase, an image $y = f(x)$ under an unknown key x is received. The goal is to find the unknown key x by making use of the precomputed tables. The main idea is to store only a part of the tables. This incurs a cost in the online phase and leads to a trade-off between the memory and time requirements.

5.1 Parallel Implementation

The pre-computation phase is essentially an exhaustive search which is required to be done only once. Practical implementations of TMTO attack will use parallel f -invocation units to perform the pre-computation. The problem that we consider is of generating the start points on chip. We show an LFSR based method for doing this. But before that, we consider the counter based method (and its disadvantage) proposed in the literature.

Counter Based Start Point Generation: Quisquater and Standaert [12] describe a generic architecture for the hardware implementation of Hellman + DP method. Nele Mentens et al [8] propose a hardware architecture for key search based on rainbow method. A global s -bit counter is used [8] as a start point generator which is connected to each of the processor. This approach has at least the following problems.

- In the analysis of success probability for TMTO, the start points are assumed to be randomly chosen. Using a counter to generate start points violates this assumption.
- Using a global s -bit counter (adder) to generate start points for n processors has the following disadvantage. Some (or all) of the n processors may ask for a start point at the same time. Then there will be a delay since there is only one global counter to generate the start points.
- On the other hand, using n counters will require n adders which can be quite expensive.

LFSR Based Start Point Generation: To generate r tables with size $m \times t$, we require a total of $m \times r$ many s -bit start points. Suppose we have n many processors P_1, P_2, \dots, P_n available for the pre-computation phase. We may assume $n|m$, since both are usually powers of two and $n < m$.

We choose n distinct primitive polynomials $p_1(x), \dots, p_n(x)$ and set-up a local start point generator (SPG) for processor P_i as follows. The local SPG is an implementation of a maximal length LFSR L_i with connection polynomial $p_i(x)$. The initial condition S_i for L_i is chosen randomly and loaded into L_i during the set-up procedure. For preparing a single table all the n processors run in parallel. For each table m/n chains need to be computed. This is done by requiring each processor to compute m/n chains. The description of P_i is as follows.

P_i : U_i denotes the current state of L_i ;

1. $U_i \leftarrow S_i; j \leftarrow 1$;
2. do while ($j \leq \frac{m}{n}$)
3. generate the *chain* with start point U_i ;
4. if the *chain* reaches an end point T_i
5. store (S_i, T_i) into Tab_i ;
6. $j \leftarrow j + 1$;
7. end if;
8. $U_i = \text{next}_i(U_i)$;
9. end do

end.

The function $\text{next}_i()$ refers to clocking LFSR L_i once. In this design, each processor P_i has its own SPG as opposed to a global SPG for all the P_i 's. This simplifies the design considerably while retaining the pseudo-random characteristic of start points. Further, as discussed earlier, implementing the LFSRs is significantly more cost effective and faster than implementing counters in hardware.

6 Application 3: Counter Mode of Operation

In 1979, Diffie and Hellman [4] introduced the counter mode (CTR mode) of operation for a block cipher. This mode actually turns a block cipher into an additive stream cipher. Let $E_k()$ be a $2s$ -bit block cipher. The pseudorandom sequence is produced as follows:

$$E_k(\text{nonce}||S_0)||E_k(\text{nonce}||S_1)||E_k(\text{nonce}||S_2)||\dots,$$

where nonce is an s -bit value and S_0, S_1, \dots is a sequence of s -bit values. The security requirements are the following.

1. The nonce is changed with each message such that the same (key,nonce) pair is never repeated.
2. The sequence S_0, S_1, S_2, \dots is a non-repeating sequence.

Usual implementations define $S_i = \text{bin}_s(i)$, where $\text{bin}_s(i)$ is the s -bit representation of the integer i . With this definition, the sequence S_i can be implemented using a counter.

Hardware implementation of CTR mode can incorporate a high degree of parallel processing. The inherent parallelism is that each $2s$ -bit block of pseudo-random bits can be produced in parallel. Suppose we have n many processors P_0, P_1, \dots, P_{n-1} where each processor is capable of one block cipher encryption. Processor P_i encrypts the values $\text{nonce}||S_i, \text{nonce}||S_{n+i}, \text{nonce}||S_{2n+i}, \dots$. If S_i is defined to be $\text{bin}_s(i)$, then there are two ways of generating the sequence.

Single adder: With a single adder, the algorithm proceeds as follows. At the start of the j th round ($j \geq 1$), the adder generates the values $S_{n(j-1)}, \dots, S_{nj-1}$. Then all the processors operate in parallel and processor P_i encrypts $\text{nonce}||S_{n(j-1)+i}$.

Problem: The single adder introduces delay which affects the overall performance of the parallel implementation.

n **adders:** In this case, each P_i has its own adder. Its local counter is initialized to S_i and after each block cipher invocation, the adder adds n to the local counter.

Problem: In this implementation, the cost of implementing n adders can take up chip area which is better utilised otherwise.

6.1 LFSR Based Solution

Note that the only restriction on the sequence S_0, S_1, \dots is that it is non-repeating. Thus, one can use a maximal length LFSR with a primitive connection polynomial to generate the sequence. Again there are two approaches to the design both of which are better than the corresponding approach based on using adders.

Single LFSR: In this case, a single LFSR is used which is initialised with a non-zero s -bit value. For $j \geq 1$, before the start of the j th round, the LFSR is clocked n times to produce the values $S_{n(j-1)}, \dots, S_{nj-1}$. P_i then encrypts $\text{nonce} \parallel S_{n(j-1)+i}$ as before. Clocking the LFSR n times introduces a delay of only n clocks into the system. This is significantly less than the time required for n increments using an adder.

n **LFSRs:** We can avoid the delay of n clocks by using n different implementations of the same LFSR initialised by suitable s -bit values to ensure that the sequences generated by the implementations are pairwise disjoint. The description of how this can be done is given in Section 3. As discussed earlier, the cost of n separate implementations of the same LFSR scales linearly with the value of n and does not consume too much chip area.

Using the LFSR based method to generate the sequence S_0, S_1, \dots will lead to an improved price-performance ratio compared to the counter based method. The design must specify the actual LFSR being used, and the required initial condition(s). Since there are many maximal length LFSRs to choose from, this provides additional flexibility to the designer.

6.2 Salsa20 Stream Cipher

Salsa20 [2] is an additive stream cipher which has been proposed as a candidate for the recent ECRYPT call for stream cipher primitives. The core design of Salsa20 consists of a hash function which is used in the counter mode to obtain a stream cipher. Denote by $\text{Salsa20}_k()$ the Salsa20 hash function. Then the pseudorandom stream is defined as follows.

$$\text{Salsa20}_k(v, S_0), \text{Salsa20}_k(v, S_1), \text{Salsa20}_k(v, S_2), \dots$$

where v is a 64-bit nonce and $S_i = \text{bin}_{64}(i)$. For hardware implementation, we can possibly generate the sequence S_0, S_1, \dots using an LFSR as described above. This defines a variant of the Salsa20 stream cipher algorithm. We believe that this modification does not diminish the security of Salsa20.

6.3 Discussion

For certain algorithms replacing counters by LFSRs will not provide substantial improvements. For example, hardware implementation of $Salsa20_k()$ will require an adder since addition operation is required by the Salsa20 algorithm itself. Hence, avoiding the adder for generating the sequence S_0, S_1, S_2, \dots might not provide substantial improvements. On the other hand, let us consider AES. No adder is required for hardware implementation of AES. Hence, using LFSR(s) to produce the sequence S_0, S_1, S_2, \dots will ensure that no adder is required for hardware implementation of the counter mode of operation. In this case, the benefits of using LFSRs will be more pronounced.

References

- [1] Electronics Frontier Foundation, Cracking DES, O'Reilly and Associates, 1998.
- [2] D. J. Bernstein. Salsa20 specification, ecrypt submission 2005. <http://www.ecrypt.eu.org/>
- [3] S. Burman and P. Sarkar. An Efficient Algorithm for Software Generation of Binary Linear Recurrences, *Appl. Algebra Eng. Commun. Comput.* 15(3-4): 201-203 (2004)
- [4] W. Diffie and M. Hellman. Privacy and Authentication: An Introduction to Cryptography, *Proceedings of the IEEE*, 67, pp. 397-427, 1979.
- [5] M. Hellman. A cryptanalytic Time-Memory Trade-off, *IEEE Transactions on Information Theory*, vol 26, pp 401-406, 1980.
- [6] R. Lidl and H. Niederreiter. Introduction to Finite Fields and their applications, Cambridge University Press, Cambridge, pp 189-249, 1994 (revised edition).
- [7] A.J. Menezes, P.C. van Oorschot and S.A. Vanstone. Handbook of Applied Cryptography, pp 195-201. CRC, Boca Raton, 2001.
- [8] N. Mentens, L. Batina, B. Preneel, and I. Verbauwhede. Cracking Unix passwords using FPGA platforms, Presented at SHARCS'05, 2005.
- [9] S. Mukhopadhyay and P. Sarkar. Application of LFSRs in Time/Memory Trade-Off Cryptanalysis, in the proceedings of WISA 2005, LNCS, to appear.
- [10] S. Mukhopadhyay and P. Sarkar. Application of LFSRs for Parallel Sequence Generation in Cryptologic Algorithms, *Cryptology ePrint Technical report 2006/042*, <http://eprint.iacr.org/2006/042>, 6 Feb, 2006.
- [11] P. Oechslin. Making a faster Cryptanalytic Time-Memory Trade-Off, in the proceedings of CRYPTO 2003, LNCS, vol 2729, pp 617-630, 2003.
- [12] J.J. Quisquater and F.X. Standaert. Exhaustive Key Search of the DES: Updates and Refinements, presented at SHARCS'05, 2005.
- [13] M.J. Wiener. Efficient DES Key Search, presented at the rump session of CRYPTO 1993, reprinted in *Practical Cryptography for Data Internetworks*, W. Stallings editor, IEEE Computer Society Press, 1996, pp. 31-79.

Provable Security for an RC6-like Structure and a MISTY-FO-like Structure Against Differential Cryptanalysis

Changhoon Lee¹, Jongsung Kim^{1,2}, Jaechul Sung³,
Seokhie Hong¹, and Sangjin Lee¹

¹ Center for Information Security Technologies(CIST),
Korea University, Anam Dong, Sungbuk Gu, Seoul, Korea
{crypto77, joshep, hsh, sangjin}@cist.korea.ac.kr

² Katholieke Universiteit Leuven, ESAT/SCD-COSIC, Belgium
Kim.Jongsung@esat.kuleuven.be

³ Department of Mathematics, University of Seoul, 90,
Cheonnong Dong, Dongdaemun Gu, Seoul, Korea
jcsung@uos.ac.kr

Abstract. In this paper we introduce two new block cipher structures, named *RC6-like structure* and *MISTY-FO-like structure*, and show that these structures are provably resistant against differential attack. The main results of this paper are that the 5-round differential probabilities of these structures are upperbounded by $p^4 + 2p^5$ and p^4 , respectively, if the maximum differential probability of a round function is p . We also discuss a provable security for the RC6-like structure against LC. Our results are attained under the assumption that all of components in our proposed structures are bijective.

Keywords: Provable Security, Differential Cryptanalysis, Linear Cryptanalysis, RC6, MISTY, Feistel Network.

1 Introduction

One of the most powerful attacks on block ciphers is Differential Cryptanalysis(DC) [2] which was introduced by Biham and Shamir. The general purpose of a differential attack is to find the first or the last round keys with a complexity less than an exhaustive search for a master key by using a characteristic with a high probability. However, the maximum characteristic probability may not guarantee a block cipher to be secure against DC even if it is sufficiently small. In order to show that a block cipher is secure against DC, we should prove the maximum differential probability is upperbounded by a small enough value. In other words, a provable security against DC can be achieved by finding the number of rounds such that the differential probabilities are upperbounded by a small enough value.

The other one of the most powerful attacks on block ciphers is Linear Cryptanalysis(LC) [9] which was introduced by Matsui. Similarly, we can prove the

Table 1. Comparison our results with Nyberg’s

Paper	Structure	Differential Probability	Rounds
Nyberg [12]	GFN	p^4	$r \geq 6$
This paper	RC6-like	$p^4 + 2p^5$	$r \geq 5$
This paper	MISTY-FO-like	p^4	$r \geq 5$

¹ Rounds : The specific number of rounds bounded by differential probability

security of a block cipher against LC by giving the number of rounds such that the linear hulls are upperbounded by a small enough value. It is well-known that the proof of linear hull probabilities is almost same as that of differential probabilities [1, 10, 11, 13].

Nyberg and Knudsen first proposed the conception of a provable security against DC and gave a provable security for a Feistel structure in 1992 [11]. In [12] Nyberg also proposed a conjecture that if a generalized Feistel network (GFN) has n parallel bijective F functions per round then the average probability of each differential over at least $3n$ rounds is less than or equal to p^{2n} where p is the maximum average differential probability of the F function. In [10] Matsui introduced a block cipher MISTY with a provable security against DC and LC as an example of such a construction. Furthermore, Sung et al. and Hong et al. [5, 6, 15] showed a provable security for a SKIPJACK-like and a SPN structures, respectively.

In this paper we show a provable security for an RC6-like structure (which is a generalized Feistel network different from Nyberg’s) and a MISTY-FO-like structure against DC. Our main results are that the maximum differential probabilities over at least 5 rounds of RC6-like and MISTY-FO-like structures are upperbounded by $p^4 + 2p^5$ and p^4 , respectively. Furthermore, we shortly discuss a provable security for the RC6-like structure against LC. Table 1 summarizes our results.

2 Preliminaries

Throughout this paper, we consider block cipher structures with a n -bit round function $F_k: \mathbb{GF}(2^n) \rightarrow \mathbb{GF}(2^n)$ where k is a round key. We also assume that a round key is generated independently and uniform randomly. For convenience, F_k is denoted by F .

Definition 1 (Differential Probability [10]). For any given $\Delta X, \Delta Y \in \mathbb{GF}(2^n)$, the differential probabilities of a round function F are defined as;

$$DP^F(\Delta X \rightarrow \Delta Y) = \frac{\#\{X \in \mathbb{GF}(2^n) \mid F(X) \oplus F(X \oplus \Delta X) = \Delta Y\}}{2^n}$$

Definition 2 (Linear Hull Probability [10]). For any given $\Gamma X, \Gamma Y \in \mathbb{GF}(2^n)$, the differential probabilities of a round function F are defined as;

$$LP^F(\Gamma Y \rightarrow \Gamma X) = \left(\frac{\#\{X \in \mathbb{GF}(2^n) \mid \Gamma X \bullet X = \Gamma Y \bullet F(X)\}}{2^{n-1}} - 1 \right)^2$$

where $\Gamma X \bullet \Gamma Y$ denotes the parity of bitwise exclusive-or of ΓX and ΓY .

Note that $DP^F(\Delta X \rightarrow \Delta Y)$ and $LP^F(\Gamma Y \rightarrow \Gamma X)$ are average probabilities over all possible keys. The following definition is useful to evaluate a resistance against DC and LC.

Definition 3 (Maximal Differential and Linear Hull Probabilities). *The maximal differential and linear hull probabilities of F are defined by*

$$DP_{max}^F = \max_{\Delta X \neq 0, \Delta Y} DP^F(\Delta X \rightarrow \Delta Y) = p$$

and

$$LP_{max}^F = \max_{\Gamma X, \Gamma Y \neq 0} LP^F(\Gamma Y \rightarrow \Gamma X) = q.$$

Theorem 1 ([10]). (i) *For any function F ,*

$$\sum_{\Delta Y} DP^F(\Delta X \rightarrow \Delta Y) = 1, \quad \sum_{\Gamma X} LP^F(\Gamma Y \rightarrow \Gamma X) = 1$$

(ii) *For any bijective function F ,*

$$\sum_{\Delta X} DP^F(\Delta X \rightarrow \Delta Y) = 1, \quad \sum_{\Gamma Y} LP^F(\Gamma Y \rightarrow \Gamma X) = 1.$$

We can easily show $DP^F(0 \rightarrow 0) = 1$ and $DP^F(\Delta X \rightarrow \Delta Y) \leq p$ if $\Delta X \neq 0$ or $\Delta Y \neq 0$. These facts are used significantly in the proof of Theorem 3. We can calculate differential and linear hull probabilities in consecutive two rounds with the following theorem.

Theorem 2 ([10]). *For any $\Delta X, \Delta Z, \Gamma X, \Gamma Z \in \mathbb{GF}(2^n)$,*

$$DP^{F_1, F_2}(\Delta X \rightarrow \Delta Z) = \sum_{\Delta Y} DP^{F_1}(\Delta X \rightarrow \Delta Y) \cdot DP^{F_2}(\Delta Y \rightarrow \Delta Z)$$

and

$$LP^{F_1, F_2}(\Gamma Z \rightarrow \Gamma X) = \sum_{\Gamma Y} LP^{F_2}(\Gamma Z \rightarrow \Gamma Y) \cdot LP^{F_1}(\Gamma Y \rightarrow \Gamma X).$$

3 Provable Security Against DC

In this Sect., we introduce an RC6-like and a MISTY-FO-like structures and show a provable security for them against DC.

3.1 On RC6-Like Structure

One round encryption of RC6-like structure, which is similar to the RC6 cipher [14], is described as follows.

$$Y_1 = F(X_3) \oplus X_4, \quad Y_2 = X_1, \quad Y_3 = F(X_1) \oplus X_2, \quad Y_4 = X_3$$

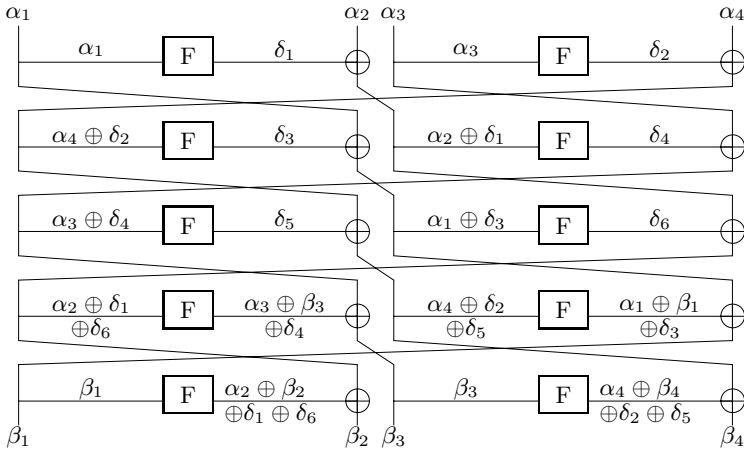


Fig. 1. Notations of 5-round differential for RC6-like Structure

Table 2. Notations used in Proof

Relations	R_1, R_2, \dots, R_k			
Variables	$\delta_{i_1, i_2, \dots, i_j}$			
step 1	A	B	C	D
step 2	E	F	G	H
step 3	I	J		

where (X_1, X_2, X_3, X_4) and (Y_1, Y_2, Y_3, Y_4) are the input and output of any round, respectively. The 5-round differential of this structure is then defined in Fig. 1.

The probability of the above 5-round differential is calculated as the following equation (1).

$$\begin{aligned}
 DP(\alpha \rightarrow \beta) = & \sum_{\delta_i, 1 \leq i \leq 6} DP(\alpha_1 \rightarrow \delta_1) \cdot DP(\alpha_3 \rightarrow \delta_2) \cdot DP(\alpha_4 \oplus \delta_2 \rightarrow \delta_3) \\
 & \cdot DP(\alpha_2 \oplus \delta_1 \rightarrow \delta_4) \cdot DP(\alpha_3 \oplus \delta_4 \rightarrow \delta_5) \cdot DP(\alpha_1 \oplus \delta_3 \rightarrow \delta_6) \quad (1) \\
 & \cdot DP(\alpha_2 \oplus \delta_1 \oplus \delta_6 \rightarrow \alpha_3 \oplus \beta_3 \oplus \delta_4) \cdot DP(\alpha_4 \oplus \delta_2 \oplus \delta_5 \rightarrow \alpha_1 \oplus \beta_1 \oplus \delta_3) \\
 & \cdot DP(\beta_1 \rightarrow \alpha_2 \oplus \beta_2 \oplus \delta_1 \oplus \delta_6) \cdot DP(\beta_3 \rightarrow \alpha_4 \oplus \beta_4 \oplus \delta_2 \oplus \delta_5)
 \end{aligned}$$

Here, we denote the 5-round differential probability by $DP(\alpha \rightarrow \beta)$, and $DP^F(\Delta X \rightarrow \Delta Y)$ by $DP(\Delta X \rightarrow \Delta Y)$.

We now explain a table used to prove Theorem 3. In Table 2, ‘‘Relations’’ means the conditions of α_i ’s, β_i ’s and δ_i ’s. ‘‘Variables’’ denotes the variables only summed over in the equation (1). The A, B, \dots, J are the probabilities as follows.

$$\begin{aligned}
 A &: DP(\alpha_1 \rightarrow \delta_1), & B &: DP(\alpha_3 \rightarrow \delta_2), & C &: DP(\alpha_4 \oplus \delta_2 \rightarrow \delta_3), \\
 D &: DP(\alpha_2 \oplus \delta_1 \rightarrow \delta_4), & E &: DP(\alpha_3 \oplus \delta_4 \rightarrow \delta_5), & F &: DP(\alpha_1 \oplus \delta_3 \rightarrow \delta_6), \\
 G &: DP(\alpha_2 \oplus \delta_1 \oplus \delta_6 \rightarrow \alpha_3 \oplus \beta_3 \oplus \delta_4), & H &: DP(\alpha_4 \oplus \delta_2 \oplus \delta_5 \rightarrow \alpha_1 \oplus \beta_1 \oplus \delta_3), \\
 I &: DP(\beta_1 \rightarrow \alpha_2 \oplus \beta_2 \oplus \delta_1 \oplus \delta_6), & J &: DP(\beta_3 \rightarrow \alpha_4 \oplus \beta_4 \oplus \delta_2 \oplus \delta_5)
 \end{aligned}$$

In steps 1, 2, 3, the factor “1” means the probability of $DP(0 \rightarrow 0)$ and the factor “ $\leq p$ ” means this probability is upperbounded by p . The factor “sum over $\delta_i(n)$ ” means the δ_i is summed over in Eq. (1) and n is the order of summing over. For instance, if a function F is bijective and β is nonzero, then $\sum_{\delta_1, \delta_2} DP(\alpha \oplus \delta_1 \rightarrow \delta_2) \cdot DP(\beta \rightarrow \delta_2) = \sum_{\delta_2} DP(\beta \rightarrow \delta_2) \cdot (\sum_{\delta_1} DP(\alpha \oplus \delta_1 \rightarrow \delta_2)) = \sum_{\delta_2} DP(\beta \rightarrow \delta_2) = 1$. The factor “ $DP(\alpha \oplus \delta_1 \rightarrow \delta_2)$ ” can be denoted as “sum over $\delta_1(1)$ ” and the other factor “ $DP(\beta \rightarrow \delta_2)$ ” as “sum over $\delta_2(2)$ ”. We call it *ordered sum*. If the number of “ $\leq p$ ” in Table 2 is \mathbf{n} , this means that $DP(\alpha \rightarrow \beta)$ is upperbounded by $p^{\mathbf{n}}$.

Theorem 3. *Differential probabilities of the r -round RC6-like structure are upperbounded by $p^4 + 2p^5$ if a function F is bijective and $r \geq 5$.*

Proof. Let $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ be an input difference and $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$ be an output difference after 5 rounds. By the assumption that a function F is bijective, we just consider $\alpha \neq 0$ and $\beta \neq 0$. Here, a case of $(\alpha_1 = 0, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 \neq 0)$ is the same as $(\alpha_1 = 0, \alpha_2 \neq 0, \alpha_3 = 0, \alpha_4 = 0)$. We call such two cases *dual*. Since this structure has various dual cases, the proof is divided by 9 cases according to input difference α . On the hand, the 5-round RC6-like structure has various impossible differentials. For instance, in case 6, we have $(\alpha_1 = 0, \alpha_2 = 0, \alpha_3 \neq 0, \alpha_4 \neq 0) \nrightarrow (\beta_1 = 0, \beta_2 = 0, \beta_3 \neq 0, \beta_4)$. We omit all these cases in our proof. We here describe a detailed analysis for case 1 and case 7-1. The rest of cases will be proved by Table 2 in Appendix A.

$$\frac{\text{Case 1. } \alpha_1 = 0, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 \neq 0}{(\text{dual: } \alpha_1 = 0, \alpha_2 \neq 0, \alpha_3 = 0, \alpha_4 = 0)}$$

In this case, we have $\delta_1 = \delta_2 = \delta_4 = \delta_5 = 0$ and $\delta_3 \neq 0, \delta_6 \neq 0$ and $\beta_3 \neq 0$. Therefore, variables δ_3 and δ_6 will be only summed over in Eq. (1) and $DP(\alpha_4 \rightarrow \delta_3), DP(\delta_6 \rightarrow \beta_3), DP(\alpha_4 \rightarrow \beta_1 \oplus \delta_3)$ and $DP(\beta_3 \rightarrow \alpha_4 \oplus \beta_4)$ are upperbounded by p . So we have

$$\begin{aligned}
 DP(\alpha \rightarrow \beta) &= \sum_{\delta_3, \delta_6} DP(\alpha_4 \rightarrow \delta_3) \cdot DP(\delta_3 \rightarrow \delta_6) \cdot DP(\delta_6 \rightarrow \beta_3) \\
 &\quad \cdot DP(\alpha_4 \rightarrow \beta_1 \oplus \delta_3) \cdot DP(\beta_1 \rightarrow \beta_2 \oplus \delta_6) \cdot DP(\beta_3 \rightarrow \alpha_4 \oplus \beta_4) \\
 &\leq p^4 \cdot \sum_{\delta_3, \delta_6} DP(\delta_3 \rightarrow \delta_6) \cdot DP(\beta_1 \rightarrow \beta_2 \oplus \delta_6) \\
 &\leq p^4 \cdot \sum_{\delta_6} DP(\beta_1 \rightarrow \beta_2 \oplus \delta_6) \cdot \left(\sum_{\delta_3} DP(\delta_3 \rightarrow \delta_6) \right) = p^4
 \end{aligned}$$

If β_1 is zero, we have $\delta_6 = \beta_2$ and $DP(\beta_1 \rightarrow \beta_2 \oplus \delta_6) = 1$. But we can regard δ_6 as a variable in the above equation though β_1 is zero. Table 3 summarizes the above equation.

Table 3. Proof of Case 1

Relations	$\delta_{1,2,4,5} = 0, \delta_{3,6} \neq 0, \beta_3 \neq 0$			
Variables	$\delta_{3,6}$			
step 1	1	1	$\leq p$	1
step 2	1	sum over $\delta_3(1)$	$\leq p$	$\leq p$
step 3	sum over $\delta_6(2)$	$\leq p$		

Case 7-1. $\alpha_1 \neq 0, \alpha_2 \neq 0, \alpha_3 = 0, \alpha_4 \neq 0$ ($\beta_3 \neq 0$)

By this assumption, we have $\delta_1 \neq 0, \delta_2 = 0$ and $\delta_3 \neq 0$. We can calculate Eq. (1) as follows.

$$\begin{aligned}
 DP(\alpha \rightarrow \beta) &= \sum_{\delta_i, \{1,3,4,5,6\}} DP(\alpha_1 \rightarrow \delta_1) \cdot DP(\alpha_4 \rightarrow \delta_3) \\
 &\quad \cdot DP(\alpha_2 \oplus \delta_1 \rightarrow \delta_4) \cdot DP(\delta_4 \rightarrow \delta_5) \cdot DP(\alpha_1 \oplus \delta_3 \rightarrow \delta_6) \\
 &\quad \cdot DP(\alpha_2 \oplus \delta_1 \oplus \delta_6 \rightarrow \beta_3 \oplus \delta_4) \cdot DP(\alpha_4 \oplus \delta_5 \rightarrow \alpha_1 \oplus \beta_1 \oplus \delta_3) \\
 &\quad \cdot DP(\beta_1 \rightarrow \alpha_2 \oplus \beta_2 \oplus \delta_1 \oplus \delta_6) \cdot DP(\beta_3 \rightarrow \alpha_4 \oplus \beta_4 \oplus \delta_5)
 \end{aligned}$$

(i) $\delta_4 \neq 0$

From this assumption, $DP(\alpha_1 \rightarrow \delta_1)$, $DP(\alpha_4 \rightarrow \delta_3)$, $DP(\alpha_2 \oplus \delta_1 \rightarrow \delta_4)$ and $DP(\delta_4 \rightarrow \delta_5)$ are upperbounded by p . So we have

$$\begin{aligned}
 DP(\alpha \rightarrow \beta) &\leq p^4 \cdot \sum_{\delta_i, \{1,3,4,5,6\}} DP(\alpha_1 \oplus \delta_3 \rightarrow \delta_6) \cdot DP(\alpha_2 \oplus \delta_1 \oplus \delta_6 \rightarrow \beta_3 \oplus \delta_4) \\
 &\quad \cdot DP(\alpha_4 \oplus \delta_5 \rightarrow \alpha_1 \oplus \beta_1 \oplus \delta_3) \cdot DP(\beta_1 \rightarrow \alpha_2 \oplus \beta_2 \oplus \delta_1 \oplus \delta_6) \\
 &\quad \cdot DP(\beta_3 \rightarrow \alpha_4 \oplus \beta_4 \oplus \delta_5) \leq p^4 \cdot \sum_{\delta_5} DP(\beta_3 \rightarrow \alpha_4 \oplus \beta_4 \oplus \delta_5) \\
 &\quad \cdot \sum_{\delta_3} DP(\alpha_4 \oplus \delta_5 \rightarrow \alpha_1 \oplus \beta_1 \oplus \delta_3) \cdot \sum_{\delta_6} DP(\alpha_1 \oplus \delta_3 \rightarrow \delta_6) \\
 &\quad \cdot \sum_{\delta_1} DP(\beta_1 \rightarrow \alpha_2 \oplus \beta_2 \oplus \delta_1 \oplus \delta_6) \cdot \sum_{\delta_4} DP(\alpha_2 \oplus \delta_1 \oplus \delta_6 \rightarrow \beta_3 \oplus \delta_4) \leq p^4.
 \end{aligned}$$

(ii) $\delta_4 = 0$

Since $\delta_4 = 0$, we have $\delta_1 = \alpha_2$ and $\delta_5 = 0$, and $DP(\alpha_1 \rightarrow \delta_1)$, $DP(\alpha_4 \rightarrow \delta_3)$, $DP(\alpha_2 \oplus \delta_1 \oplus \delta_6 \rightarrow \beta_3 \oplus \delta_4)$, $DP(\alpha_4 \oplus \delta_5 \rightarrow \alpha_1 \oplus \beta_1 \oplus \delta_3)$ and $DP(\beta_3 \rightarrow \alpha_4 \oplus \beta_4 \oplus \delta_5)$ are also upperbounded by p . So we have

$$\begin{aligned}
 DP(\alpha \rightarrow \beta) &= \sum_{\delta_3, \delta_6} DP(\alpha_1 \rightarrow \alpha_2) \cdot DP(\alpha_4 \rightarrow \delta_3) \cdot DP(\alpha_1 \oplus \delta_3 \rightarrow \delta_6) \cdot DP(\delta_6 \rightarrow \beta_3) \\
 &\quad \cdot DP(\alpha_4 \rightarrow \alpha_1 \oplus \beta_1 \oplus \delta_3) \cdot DP(\beta_1 \rightarrow \beta_2 \oplus \delta_6) \cdot DP(\beta_3 \rightarrow \alpha_4 \oplus \beta_4) \\
 &\leq p^5 \cdot \sum_{\delta_6} DP(\beta_1 \rightarrow \beta_2 \oplus \delta_6) \cdot \left(\sum_{\delta_3} DP(\alpha_1 \oplus \delta_3 \rightarrow \delta_6) \right) \leq p^5.
 \end{aligned}$$

Therefore, the differential probability of this case 7-1 is upperbounded by $p^4 + p^5$. This proof is summarized in Table 4.

Table 4. Proof of Case 7-1: ($\beta_3 \neq 0$)

Relations	$\delta_{1,3} \neq 0, (*\delta_4 \neq 0)$			
Variables	$\delta_{1,3,4,5,6}$			
step 1	$\leq p$	1	$\leq p$	$\leq p$
step 2	$\leq p$	sum over $\delta_6(3)$	sum over $\delta_4(1)$	sum over $\delta_3(4)$
step 3	sum over $\delta_1(2)$	sum over $\delta_5(5)$		
Relations	$\delta_{1,3} \neq 0, \delta_1 = \alpha_2, \delta_5 = 0 (*\delta_4 = 0)$			
Variables	$\delta_{3,6}$			
step 1	$\leq p$	1	$\leq p$	1
step 2	1	sum over $\delta_3(1)$	$\leq p$	$\leq p$
step 3	sum over $\delta_6(2)$	$\leq p$		

For want of space, we can't present the proofs of other cases in the paper. If you want to see detailed proofs then you can find them in the full version of the paper [8]. The appendix A of [8] shows that the differential probabilities of cases 8-3, 9-1, 9-2, 9-3 are upperbounded by $p^4 + 2p^5$, and that the differential probabilities of the rest cases are upperbounded by p^4 or $p^4 + p^5$, respectively. Hence, the maximal differential probabilities of this 5-round structure are less than or equal to $p^4 + 2p^5$. It follows that r -round differential probabilities are upperbounded by $p^4 + 2p^5$, if $r \geq 5$. \square

Since the RC6-like structure can be regarded as one of the generalizations of Feistel structure, a provable security against LC can be also obtained as in [1, 10, 11, 13]. Indeed, the linear hull propagations for RC6-like structure are exactly the same as the differential propagations for it, so the linear hull probabilities for the r -round RC6-like structure are upperbounded by $q^4 + 2q^5$ where q is the maximum linear hull probability of a function F , if F is bijective and $r \geq 5$.

3.2 On MISTY-FO-like Structure

We introduce a MISTY-FO-like structure which is similar to the round function FO of the original MISTY block cipher and show a provable security for this sturture against DC. One round encryption of this structure is formally defined as follows (see Fig. 2).

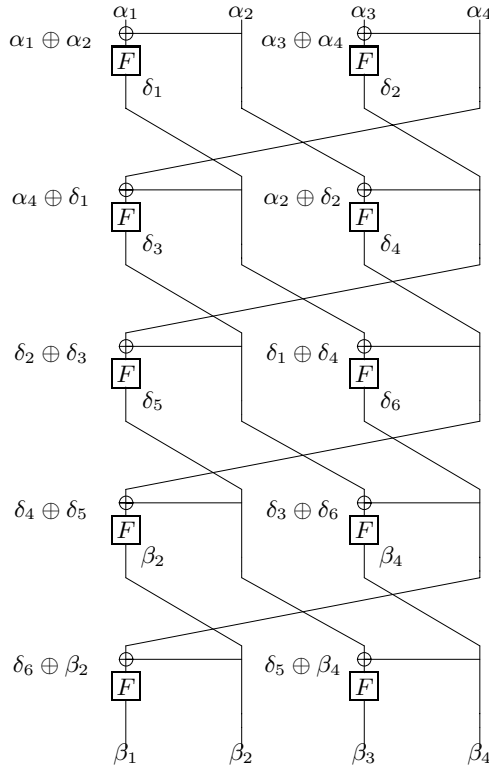


Fig. 2. Notation of 5-round Differential for MISTY-FO-like Structure

$$Y_1 = X_4, Y_2 = F(X_1 \oplus X_2), Y_3 = X_2, Y_4 = F(X_3 \oplus X_4)$$

where (X_1, X_2, X_3, X_4) and (Y_1, Y_2, Y_3, Y_4) are the input and output of any round, respectively, and F is a keyed round function.

We can then represent the 5-round differentials of this structure as in Fig. 2, and calculate the 5-round differential probabilities as the following equation.

$$\begin{aligned}
 DP(\alpha \rightarrow \beta) &= \sum_{\delta_i, 1 \leq i \leq 6} DP(\alpha_1 \oplus \alpha_2 \rightarrow \delta_1) \\
 &\cdot DP(\alpha_3 \oplus \alpha_4 \rightarrow \delta_2) \cdot DP(\alpha_4 \oplus \delta_1 \rightarrow \delta_3) \\
 &\cdot DP(\alpha_2 \oplus \delta_2 \rightarrow \delta_4) \cdot DP(\delta_2 \oplus \delta_3 \rightarrow \delta_5) \\
 &\cdot DP(\delta_1 \oplus \delta_4 \rightarrow \delta_6) \cdot DP(\delta_4 \oplus \delta_5 \rightarrow \beta_2) \\
 &\cdot DP(\delta_3 \oplus \delta_6 \rightarrow \beta_4) \cdot DP(\delta_6 \oplus \beta_2 \rightarrow \beta_1) \\
 &\cdot DP(\delta_5 \oplus \beta_4 \rightarrow \beta_3)
 \end{aligned}$$

Theorem 4. *Differential probabilities of the r -round MISTY-FO-like structures are upperbounded by p^4 if a function F is bijective and $r \geq 5$.*

Since the proof skill is similar to that of Theorem 3 we also omit a concrete proof of this theorem. See the full version of the paper [8] for our detailed proofs.

4 Conclusion

In this paper we have studied a provable security of two new block cipher structures named a RC6-like and a MISTY-FO-like structures against DC. In order to estimate their security against DC, we have used the upper bound of differential probability. Table 1 shows that the differential probabilities over at least 5 rounds for an RC6-like and a MISTY-FO-like structures are less than or equal to $p^4 + 2p^5$ and p^4 , respectively. Furthermore, the linear hull probabilities over at least 5 rounds for the RC6-like structure are less than or equal to $q^4 + 2q^5$. We believe that these results are very helpful to design provably secure block ciphers against DC and LC.

Acknowledgments

We would like to thank the anonymous referees for helpful comments about this work. This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment). The second author was financed by a Ph.D. grant of the Katholieke Universiteit Leuven and by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD) (KRF-2005-213-D00077) and supported by the Concerted Research Action (GOA) Ambiorics 2005/11 of the Flemish Government and by the European Commission through the IST Programme under Contract IST2002507932 ECRYPT.

References

1. K. Aoki and K. Ohta, *Strict Evaluation of the Maximum Average of Differential Probability and the Maximem Average of Linear Probability*, IEICE Transcations fundamentals of Elections, Communications and Computer Sciences, No. 1, pp. 2–8, 1997.
2. E. Biham and A. Shamir, *Differential cryptanalysis of DES-like cryptosystems*, Advances in Cryptology – CRYPTO’90, LNCS 537, Springer-Verlag, 1991, pp. 2–21.
3. E. Biham and A. Shamir, *Differential cryptanalysis of the full 16-round DES*, Advances in Cryptology – CRYPTO’92, LNCS 740, pp. 487–496, Springer-Verlag, 1992.
4. J. Daemen and V. Rijndael, *The Rijndael block cipher*, AES proposal, 1998.
5. S. Hong, S. Lee, J. Lim, J. Sung, D. Cheon, and I. Cho, *Provable Security against Differential and Linear Cryptanalysis for the SPN Structure*, FSE’00, LNCS 1978, pp. 273–283, Springer-Verlag, 2001.
6. S. Hong, J. Sung, S. Lee, J. Lim, and J. Kim, *Provable Security for 13 round Skipjack-like Structure*, Information Processing Letters, 2001.

7. M. Kanda, Y. Takashima, T. Matsumoto, K. Aoki, and K. Ohta, *A strategy for constructing fast functions with practical security against differential and linear cryptanalysis*, SAC'99, LNCS 1556, pp. 264–279, Springer-Verlag, 1999.
8. C. Lee, J. Kim, J. Sung, S. Hong, and S. Lee, *Provable Security for an RC6-like Structure and a MISTY-FO-like Structure against Differential Cryptanalysis-Full Version*. Available at <http://homes.esat.kuleuven.be/~kjongsun/publication.html> (or <http://cist.korea.ac.kr/new/Publication/index.html>).
9. M. Matsui, *Linear cryptanalysis method for DES cipher*, Advances in Cryptology – EUROCRYPT'93, LNCS 765, pp. 386–397, Springer-Verlag, 1994.
10. M. Matsui, *New structure of block ciphers with provable security against differential and linear cryptanalysis*, FSE'96, LNCS 1039, pp. 205–218, Springer-Verlag, 1996.
11. K. Nyberg and L.R. Knudsen, *Provable security against differential cryptanalysis*, Advances in Cryptology – CRYPTO'92, LNCS 740, pp. 566–574, Springer-Verlag, 1992.
12. K. Nyberg *Generalized Feistel Networks*, Advances in Cryptology – ASIACRYPT'96, LNCS 1163, pp. 91–104, Springer-Verlag, 1996.
13. K. Nyberg, *Linear approximation of block ciphers*, Presented at rump session, Eurocrypt'94, May 1994.
14. R. Rivest, M. Robshaw, R. Sidney, and Y.L. Yin, *The RC6 Block Cipher*, <http://theory.lcs.mit.edu/~rivest/rc6.pdf>.
15. J. Sung, S. Lee, J. Lim, S. Hong, and S. Park, *Provable Security for the Skipjack-like Structure against Differential Cryptanalysis and Linear Cryptanalysis*, Advances in Cryptology – ASIACRYPT'00, LNCS 1976, pp. 274–288, Springer-Verlag, 2000.

Design and Implementation of an FPGA-Based 1.452-Gbps Non-pipelined AES Architecture

Ignacio Algreto-Badillo, Claudia Feregrino-Uribe, and René Cumplido

National Institute for Astrophysics, Optics and Electronics,
Luis Enrique Erro #1, CP 72840, Sta. Ma. Tonantzintla, Puebla, México
{algreodobadillo, cferegrino, rcumplido}@inaoep.mx
<http://ccc.inaoep.mx>

Abstract. This work reports a non-pipelined AES (Advanced Encrypted Standard) FPGA (Field Programmable Gate Array) architecture, with low resource requirements. The architecture is designed to work on CBC (Cipher Block Chaining) mode and achieves a throughput of 1.45 Gbps. This implementation is a module of a configuration library for a Cryptographic Reconfigurable Platform (CRP).

1 Introduction

Cryptographic systems in diverse applications, like multimedia, WWW servers, the Transport Layer Security (TLS) protocol and secure mail protocols such as S/MIME, have provided a safe way for storing and transmitting information. These systems offer security based on complex architectures by adding cryptographic algorithms that may be hash functions, symmetric key algorithms and asymmetric key algorithms [1]. Each one can be used for multiple and different services, such as: authentication, integrity, data confidentiality, pseudorandom number generator, digital signatures and key establishment.

Secure communication protocols handle a symmetric key cryptosystem, a hash algorithm, and method for providing digital signatures and key exchange using public key cryptography [2] and have several operation modes. This work focuses on IPsec (Internet Protocol Security) due to its growing popularity [3], it operates at the network IP layer of the TCP/IP stack and utilizes CBC mode. It is an algorithm-independent protocol, securing traffic of whichever network topology. It has a set of open standards and protocols for creating and maintaining secure communications over IP networks. For example, IPsec leverages other important standards like IKE (Internet Key Exchange) protocol, authentication standards (Kerberos, RADIUS, and X.509 digital certificates), and encryption algorithms (AES and 3DES). Communication networks, like Gigabit Ethernet require processing speeds of 1 Gbps and it is expected that also future wireless personal area networks perform at these data rates [4]. These networks require flexible, high throughput systems which compute cryptographic algorithms that are more efficiently implemented in custom hardware than in software running on general-purpose processors (GPPs) [5]. Also, the hardware implementations offer more security than software ones because they cannot be as easily read or modified by an outside attacker [6]. Implementing cryptography in FPGA

devices provides a good alternative to custom and semi custom ASICs (Application Specific Integrated Circuits), which have an expensive and time-consuming fabrication, and more inflexibility or parameter switching [7], and GPPs and special-purpose processors, like DSPs (Digital Signal Processors) [8], that offer lower performance. The advantages of the FPGA reprogrammable devices are especially prominent in security applications, where new security protocols decouple the choice of cryptographic algorithms from the design of the protocol, and users select the cryptographic standard to start a secure session.

AES algorithm ensures the compliance with many of the security regulations, including IPsec and Suite B of recommended cryptographic algorithms, with keys sizes of 128 and 256 bits as announced by NSA (National Security Agency) at the 2005 RSA Conference [9].

2 AES Algorithm

The AES algorithm is a symmetric block cipher that can process data blocks of 128 bits, and it uses cipher keys of 128, 192, and 256 bits [10]. This work implements a hardware architecture of the AES algorithm for ciphering 128-bit data with 128-bit keys. Key length of 128 bits is selected, because it fits in the current IKE, and has the benefits of performance. One initial round is performed and ten round functions (see Fig. 1), where a round function has four transformations (non-linear byte substitution, constant Galois field multiplication, key addition, and an S-box substitution) to obtain the cipher text. An important operation is key expansion, which computes a key schedule or a 128-bits key in each round. The non-linear byte substitution and key expansion operations require S-box substitution, where one byte is substituted and determined by the intersection of the row and the column. These substitution values for the byte xy are defined in [10].

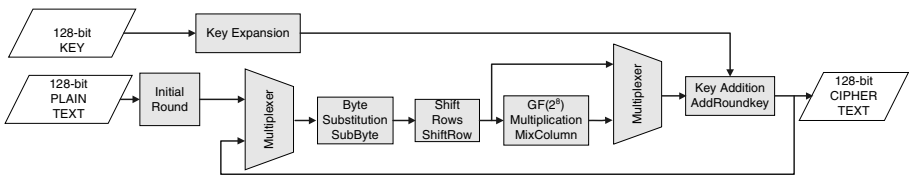


Fig. 1. General structure of AES Algorithm

3 Related Work

Recently, several algorithm-specific hardware architectures of AES algorithm have been reported in the literature, involving commercial and research works, with different design techniques, architectures and FPGA resources. Among these are iterative architectures on Virtex-II, Virtex-4, and a pipelined architecture on Virtex-II.

The commercial implementation in [11] has an iterative architecture, 128-bit data input, data output and key input buses. The datasheet presents FPGA implementations, where the “Fast version” has a throughput of 0.58 Gbps. The work in [12] presents a partitioning architecture without using the BRAMs (Blocks RAM). The

architecture is designed in two parts: 1) implementation of the Key Expansion, which calculates the round keys, and 2) implementation of the functional rounds to cipher 128-bit data. The implementation results of the second part show a throughput of 0.20 Gbps. In [13] AES implementation synthesis results are reported with three different key lengths, and the best throughput is 1.19 Gbps with 128-bit data buses. [14] describes an AES commercial product, which offers diverse operation modes and key lengths. The architecture uses 4 BRAMs in CBC mode to cipher data; the implementation needs 44 clock cycles at 93 MHz, performing at 0.27 Gbps.

The next two works are included for comparing their throughputs and FPGA resource utilization. The work in [15] reports four AES pipelined architectures, where two of them use BRAMs. The 7-stage AES architecture shows the highest throughput of 21.64 Gbps, at the expense of FPGA resources. [16] presents commercial implementations on the Xilinx Virtex-II FPGA, the main characteristics are a throughput of about 1.40 Gbps, using 18 BRAMs and 1,125 slices.

The previous works demonstrate that implementing S-box on internal memory improves the throughput, decreases the used FPGA resources, and reduces the critical path time. Current architectures with greater throughputs use pipelined structures, which are mentioned only as a reference.

This work reports an AES architecture that aims to perform above 1 Gbps to meet the speed requirements of a Cryptographic Reconfigurable Platform that changes its configuration and functionality to suit the required cryptographic implementation.

4 AES Hardware Architecture and FPGA Implementation

In a communication line or in a transmission channel, the CBC operation mode does not permit pipelined architectures, because feedback operations are performed after ciphering a block [17] (see Fig. 2).

The architecture implemented is based on the AES standard algorithm specified in the Federal Information Processing Standards Publication 197 (FIPS-197) [10] of the National Institute of Standards and Technology. It performs in CBC mode to meet the IPsec requirements.

The aim of this work is to implement a fast and simple iterative AES architecture with low FPGA resource requirements. It was written and simulated in Active-HDL and implemented in Xilinx ISE 6 for the measurement of hardware parameters such as logic and clock frequency. The architectures were synthesized, mapped, placed and routed for an FPGA Xilinx XC2V1000-FG456. This device validates the architecture, and by no means is for a final product.

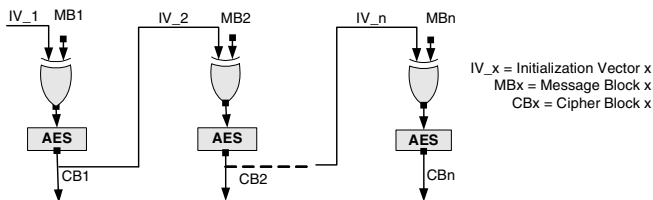


Fig. 2. AES algorithm in the CBC operation mode

The main modules of the architecture are: 1) *AES_CONTROL*, which outputs control signals and organizes the dataflow, 2) *AES_GENKEY*, which outputs the round keys, and 3) *AES_ROUND*, which ciphers the data (see Fig. 3).

The initial round is computed by the *X01* gate, and the following ten rounds are executed by the *AES_ROUND* module. The round keys are added in *AES_ROUND* module and the intermediate cipher data are feedback to the same module until the final cipher data are obtained. The selection of the initial round data and the intermediate cipher data is made by the *M01* multiplexer. After several clock cycles, the final cipher data are addressed from multiplexer output.

AES_ROUND is the main module, it covers the four transformations defined in [10] (see Fig. 4). This module calculates the ten round functions, whereas the initial round operation and the key generation are externally operated.

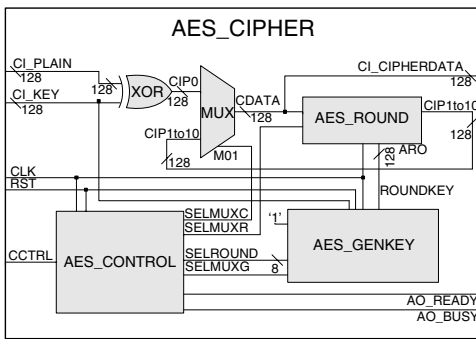


Fig. 3. General architecture of the AES implementation

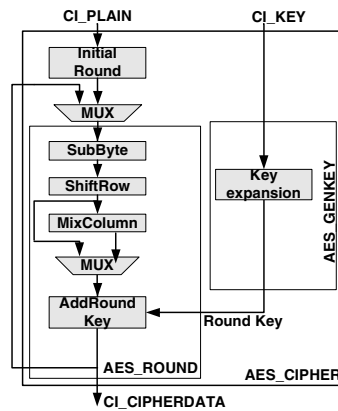


Fig. 4. The four transformations of the AES algorithm are integrated on the *AES_ROUND* module of the general architecture

The general architecture of the basic modular implementation is iterative, and the S-boxes are implemented using twenty distributed memories. *AES_Control* module is a 12-state FSM (Finite State Machine). The state diagram and FSM initial values are shown in Fig. 5. The *START* state modifies the value of the *SC_BUSY* signal. Only in this state, the signal will have value of logical zero, indicating the system is waiting for data. *SC_BUSY='1'* in other states indicates the system is busy ciphering. When *IC_STARTCIP='1'* and the actual state is *START*, FSM changes to the *LOAD* state, which registers the key and input data. In this state, the *OC_SELMUXC* signal controls the dataflow, both input data and intermediate cipher data ('1' for input data and '0' for cipher data), while *OC_SELMUXG* selects the input key or round keys ('1' for round keys and '0' for input key). Also, the initial round is computed, and the next ten states compute the ten rounds left. In each of the following clock cycles, from *ROUND1* to *ROUND10* states are active, and the *OC_ROUND* value changes for the *AES_GENKEY* module. The *ROUND10* state presents the final cipher data in the *CI_CIPHERDATA* bus. Only in this state, *OC_READY='1'* indicates a valid output. If

there are more data to cipher, the *IC_STARTCIP* signal should be in high level and the next state is *ROUND1*, else FSM is set to *START* state. In the first case, the next plain data are stored in the *ROUND10* state, and they are processed from *ROUND1* to *ROUND10* states.

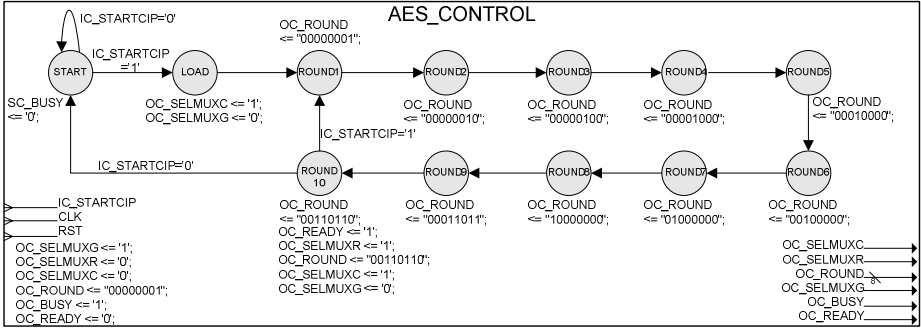


Fig. 5. State diagram of the *AES_CONTROL* module

If the system ciphers data, and it is maintained in the *ROUND0-ROUND10* loop, its output value will offer 128-bit cipher data every twelve clock cycles for 128-bit plain data and 128-bit key data. The throughput of the iterative architecture is given by [18]:

$$\text{Throughput} = \text{Plain_data_block_size} / ((\text{Clock_period})(\text{Clock_cycles})) \tag{1}$$

AES_GENKEY module is the key-expansion operation, which outputs a 128-bits key every round (see Fig. 6). S-boxes and XOR gates compute the round keys, the register stores the *CI_KEY* input or Round Key bus, and the multiplexer selects these keys. The S-boxes are implemented in four distributed memories. In the *LOAD* and *ROUND10* states the key input is stored, and from *ROUND1* to *ROUND9* states, round keys are stored. In the *ROUND10* state, the key input is stored because it is used by the *ROUND1-ROUND10* loop, when the system ciphers data successively.

The general structure of the *AES_ROUND* module is shown in the Fig. 7. This module computes the four transformations defined in [10]. The SubByte transformation is performed by S-boxes implemented in 16 distributed memories. The ShiftRow transformation is made by readdressing the *BYTESUB* bus to the *SHIFTRROW* bus. This has the effect of cyclically shifting over different number of bytes. The MixColumn transformation operates $GF(2^8)$ multiplications over *SHIFTRROW* bus, and it is performed by the *MIXCOL* module, which outputs the *MIXCOLUMN* bus. Finally, in the AddRoundKey transformation, the round key is added by a simple XOR operation. The multiplexer selects in the first nine rounds the *MIXCOLUMN* bus, whereas in the last round it selects the *SHIFTRROW* bus. The multiplexer output is added to the *IKEY* bus (round key).

The *MIXCOL* module computes multiplications and additions over $GF(2^8)$. In [10] it is described a matrix multiplication with the fixed polynomial:

$$a(x) = \{3\}x^3 + \{1\}x^2 + \{1\}x + \{2\} \tag{2}$$

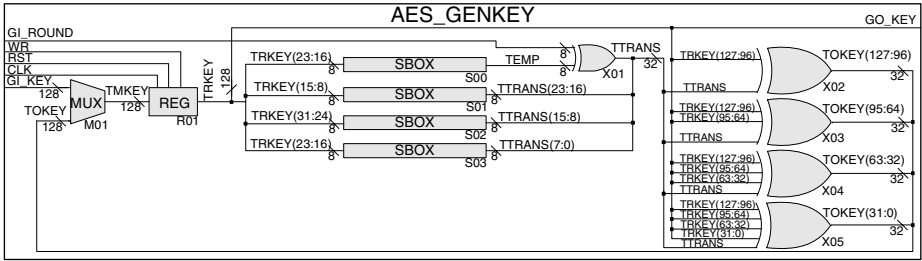


Fig. 6. Diagram of the *AES_GENKEY* module

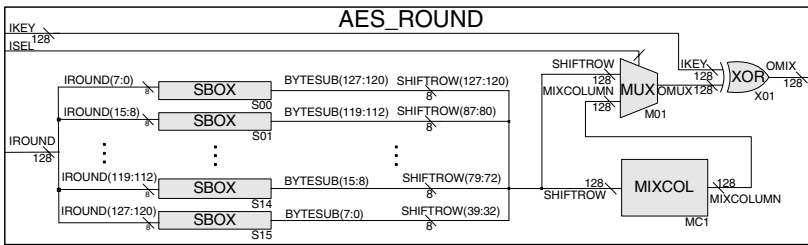


Fig. 7. Diagram of the *AES_ROUND* module

The equation computes multiplications {1} and {3}, and additions. The $GF(2^8)$ addition is the XOR operation and the $GF(2^8)$ multiplication is special since it is only necessary to multiply by some constants [19]. Constant multiplicands permit to implement XOR operations and multiplexers, and these substitute the multiplication described in [10]. For example, a section of *MIXCOL* module is shown in Fig. 8, where a MixColumn transformation is performed for the *OMIX*(127:120) byte, or

$$OMIX(127:120) \leftarrow \{2\} * IMIX(127:120) \oplus \{3\} * IMIX(119:112) \oplus \{1\} * IMIX(111:104) \oplus \{1\} * IMIX(103:96). \quad (3)$$

The {1}, {2} and {3} constant coefficients in (3) are multiplied in $GF(2^8)$. In multiplication by {1}, the result is equal to the non-one factor, so, *IMIX*(111:104) and *IMIX*(103:96) bytes are added by the *XOR49* gate.

The multiplication by {2} is a conditional 1-bit left shift implemented by a multiplexer. Its selector, *IMIX*(127), controls the overflow in $GF(2^8)$. If the value being multiplied is less than “10000000”, the result is the value itself left-shifted by 1 bit, *IMIX*(126:120)&’0’. If the value is greater than or equal to “10000000”, the result is the value left-shifted by 1 bit added with “00011011”, *IMIX*(126:120)&’0’ XOR “00011011”. This prevents overflowing and keeps the product of multiplication in $GF(2^8)$.

The multiplication by {3} is reduced to additions and multiplications by {2}, where the last multiplications are conditional 1-bit left-shifts [19]. Multiplication by {3} can be decomposed as {3} = {2} + {1}. Thus:

$$\begin{aligned}
 \{3\} * IMIX(119:112) &\leq \{2+1\} * IMIX\{119:112\} \\
 &\leq \{2\} * IMIX(119:112) + \{1\} * IMIX(119:112)
 \end{aligned}
 \tag{4}$$

The multiplication by {3} is implemented by two multiplexers, three XOR gates, and multiplications by {1} and {2} implemented as mentioned in the above paragraph.

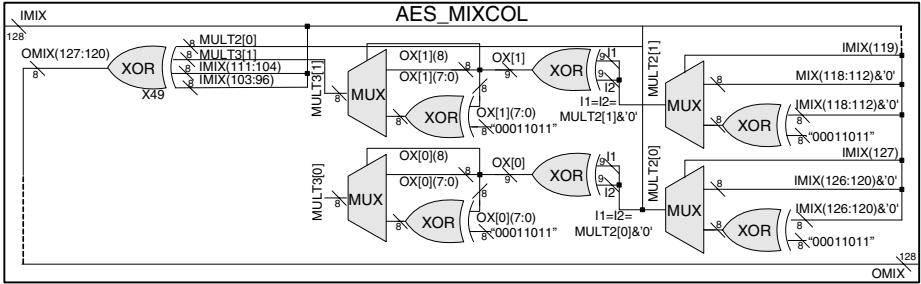


Fig. 8. Diagram of the operation in (3), which is part of the AES_ROUND module

As the IOBs requirements of the general AES architecture (see Fig. 3), exceeds the IOBs of XC2V1000-FG456 device, input data and key are stored in registers to be processed later on in parallel (see Fig. 9). So, the final architecture multiplexes the *CI_PLAIN* and *CI_KEY* 128-bit buses, requiring not additional clock cycles, since the data are stored in the last block processing time. In a given clock cycle, a bus is registered, and in the next clock cycle, the other bus. By successively ciphering data, the key and plain data are stored in run time, and each ten clock cycles, an *AO_CIP* output or cipher data are obtained.

Initially, the twenty S-boxes were implemented in twenty distributed memories, and the architecture achieved a throughput of 0.92 Gbps, with a clock frequency of 86.94 MHz, 2,335 used slices, 4,327 LUTs, 263 IOBs and 10 clock cycles.

AES architecture is part of the CRP[20], which requires cryptographic implementations for transmission speeds of 1 Gbps, and due to (1), *Plain_data_block_size* has a constant value of 128 bits, whereas *Clock_cycles* in this design has a value of 10 and *Clock_period* of $(86.94 \text{ MHz})^{-1}$. Consequently, a reduction of *Clock_cycles* implies an architecture with unrolled rounds, which increases the use of FPGA resources and the critical path time or decrease the clock frequency. A reduction of critical path time, by the modification of *Clock_period* is the more practical option.

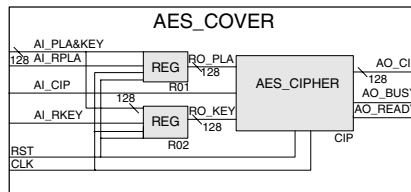


Fig. 9. Final general AES architecture

The implementation of twenty distributed memories for twenty S-boxes requires proportional FPGA resources to place and route them, which results in a critical path time proportional to the FPGA utilized logic. So, modular architecture is redesigned to use 10 dual-port embedded memories, instead of twenty, decreasing the used FPGA resources, and thus the critical path time.

5 Implementation Results and Comparisons

The implementation results of the AES architecture are shown in Table 1, and these are taken from the post-Place & Route reports. Changing the twenty internal memories by ten dual-port memories decreases the critical path time from 11.50 ns to 8.80 ns, reduces the FPGA resources, and eliminates some intermediate registers. The implementation results indicate that, in terms of required FPGA resources, the S-box substitution is the dominant element of the AES implementation.

Table 1. Implementation results of the non-pipelined iterative AES architecture

Period (ns)	Clock (MHz)	IOBs (out of 324)	Slices	4-Input LUT	Latency (Clk cycles)	Throughput (Gbps)
8.80	96.42	263	586	847	10	1.45

These results show less wired and logic FPGA resources, which are 586 slices and 847 LUTs. This leads to a compact architecture with a lower critical path time, a higher clock frequency (96.42 MHz) and a throughput of 1.45 Gbps.

Different device families will yield different performance results. Important measurements of hardware AES implementations to consider are FPGA utilized resources, clock frequency, latency and throughput, see Table 2.

Table 2 reports measurements on non-pipelined hardware architectures on Virtex and Virtex-II FPGAs, suitable for CBC mode implementation. This work reports an AES architecture with the excellent performance and low resource requirements.

Table 2. Result comparison of the AES implementations

Work / Device	FPGA Resources		Clock (MHz)	Latency (Clk cycles)	Throughput (Gbps)
	Logic	Memory			
[12] – XCV200E-6	425 CLBs	-	77.80	-	0.20
[14] – XCV250-5	791 slices	4 BRAM	93	44	-
[11] – XCV400e-8	1672 LUT	-	50.20	11	0.58
[13] – Virtex-II Pro	2703 LUT	44 BRAM	196	-	1.19
[16] – Virtex-II	1125 slices	18 BRAM	-	-	1.40
This work XC2V1000	586 slices	10 BRAM	96.42	10	1.45

The general approach used in this work aimed to obtain an iterative architecture with low hardware resources utilization. The modular design was optimized in a way that the algorithm functionality was not altered (e.g. eliminating basic modules like

registers or multiplexers). Distributed memories were replaced by dual-port memories to handle data in parallel and registers were added for data multiplexing and key storage in order to reduce the critical path, resulting in less hardware that in turn results in a more efficient place & route process and higher throughput.

6 Conclusions

This work reports a hardware architecture of AES algorithm in CBC operation mode. In terms of area requirements, throughput and hardware efficiency, this architecture exhibits excellent abilities compared to the most recent AES architectures, implemented in Virtex and Virtex II devices. This work shows a simple design and an efficient architecture that requires minimal logical resources and is suitable for devices with limited silicon area.

Its performance results and low resource requirements make the architecture suitable as a module for the CRP platform, which handles several cryptographic algorithms and is applicable in secure communication systems, where devices or networks require cryptographic solutions with high flexibility and high throughput.

References

1. P. Kocher, R. Lee, G. McGraw, A. Raghunathan, S. Ravi, Security as a New Dimension in Embedded System Design, DAC 2004, June 2004.
2. F. Crowe, A. Daly, T. Kerins, W. Marnane, Single-Chip FPGA Implementation of a Cryptographic Co-Processor, Proceedings of IEEE International Conference on Field Programmable Technology (FPT'04), December 2004.
3. Ixia, IPsec Virtual Private Networks: Conformance and Performance Testing, Whitepaper, November 2003.
4. L. Quinn, P. Mehta, A. Sicher, Wireless Communications Technology Landscape, White Paper, Dell, February 2005.
5. G. Umamaheshwari, A. Shanmugan, "Efficient VLSI Implementation of the Block Cipher Rijndael Algorithm, Academic Open Internet Journal, Volume 12, 2004. Available at: <http://www.acadjournal.com/>.
6. G. Bertoni, J. Guajardo, C. Paar, Architectures for Advanced Cryptographic Systems, Idea Group Inc, 2004.
7. K. Gaj, P. Chodowicz, Comparison of the Hardware Performance of the AES Candidates Using Reconfigurable Hardware, Proceedings of the 3rd Advanced Encryption Standard (AES) Candidate Conference, April 2000.
8. Y. Li, T. Callahan, E. Darnell, R. Harr, U. Kurkure, J. Stockwood, Hardware-Software Co-Design of Embedded Reconfigurable Architectures, ACM, 2000.
9. National Security Agency, Fact Sheet NSA Suite B Cryptography. Available at: http://www.nsa.gov/ia/industry/crypto_suite_b.cfm
10. Federal Information Processing Standards Publication 197, Announcing the Advanced Encryption Standard (AES), November 2001.
11. Barco-Silex, AES Encryption and Decryption BA411AES Factsheet, March 2005. Available at: <http://www.barco.com/>.

12. T. Liu T, C. Tanougast, P. Brunet, Y. Berviller, H. Rabah, S. Weber, An Optimized FPGA Implementation of an AES Algorithm for Embedded Applications, International Workshop on Applied Reconfigurable Computing 2005 (ARC2005), February 2005.
13. J. Lu, J. Lockwood, "IPSec Implementation on Xilinx Virtex-II Pro FPGA and Its Application", Reconfigurables Architectures Workshop (RAW), April 2005.
14. Algotronix Ltd, "AES Core Product Description", November 2004. Available at: <http://www.algotronix.com/>.
15. A. Hodjat, I. Verbauwhede, "A 21.54 Gbits/s Fully Pipelined AES Processor on FPGA", IEEE Symposium on Field-Programmable Custom Computing Machines, April, 2004.
16. Helion Technology Limited, "OVERVIEW DATASHEET – Helion cores. Available at: www.heliontech.com/.
17. A. Menezes, P. V. Oorschot, and S. Vanstone, Handbook of Applied Cryptography, CRC Press, 1996.
18. K. Gaj, P. Chodowicz, Fast Implementation and Fair Comparison of the Final Candidates for Advanced Encryption Standard Using Field Programmable Gate Array, Proceedings in RSA Security Conference – Cryptographer's Track, April 2001.
19. J. McCaffrey, You're your Data Secure with the New Advanced Encryption Standard, MSDN Magazine, Issue November 2003, Available at: <http://msdn.microsoft.com>.
20. Ignacio Algreto Badillo, René Cumplido Parra and Claudia Feregrino, "Design and Implementation of a High Performance Cryptographic Reconfigurable Platform", 2004 XIII International Congress on Computing, 13-15 Oct, Mexico (In Spanish).

Security Weaknesses in Two Proxy Signature Schemes^{*}

Jiqiang Lu

Information Security Group, Royal Holloway, University of London
Egham, Surrey TW20 0EX, U.K
Jiqiang.Lu@rhul.ac.uk

Abstract. Allowing a proxy signer to generate a signature on behalf of an original signer, a proxy signature should satisfy the property of strong unforgeability: anyone except the designated proxy signer cannot create a valid proxy signature on behalf of the original signer. Since proxy signatures, as well as their derivatives, can be used in many applications in reality, such as secure mobile agent, e-commerce systems and etc., they have been receiving extensive research recently. In this paper, we show that the proxy signature scheme [14] from ISPA'04 will suffer from the original signer's forgery attack if the original signer once gets a valid proxy signature on a message, and a similar attack arises in the proxy signature scheme [1] from AWCC'04 if the verifier does not check the originality of the proxy signer's proxy public key before verifying a proxy signature. Therefore, in some degree, neither of these two schemes meets the property of strong unforgeability.

Keywords: Public key cryptology, Proxy signature, Forgery attack.

1 Introduction

Introduced by Mambo *et al.* [7, 8], proxy signature is a new cryptographic primitive, during which an original signer firstly delegates his or her signing capability to a proxy signer and then the proxy signer creates a signature on behalf of the original signer. Based on delegation type, Mambo *et al.* classified proxy signatures as full delegation, partial delegation and delegation by warrant. In the full delegation, the original signer gives his secret key to the proxy signer. In the partial delegation, the original signer generates a proxy signature key by using his secret key and gives it to the proxy signer who uses the proxy key to sign in the following. In the delegation by warrant, the proxy signer first obtains the warrant, which is a certificate composed of a message part and a public signature key from the original signer, and then uses the corresponding secret key to sign. The resulting signature consists of the created signature and the warrant. Then, Zhang [15] proposed threshold proxy signature with the property of nonrepudiability that means neither the original signer nor the proxy signer

^{*} The work as well as the author was supported by a Royal Holloway Scholarship.

can repudiate his participation in the signature. Later, Lee *et al.* [4, 5] provided new classifications of proxy signatures as strong vs. weak proxy signatures, designated vs. non-designated proxy signatures and self-proxy signatures. Strong proxy signature represents both original signer's and proxy signer's signatures, and a proxy signer cannot repudiate his signature creation against anyone after he creates a valid proxy signature. Weak proxy signature represents only original signer's signatures and does not provide the non-repudiation of proxy signer. However, Sun *et al.* [10] presented a kind of forgery attack on Lee *et al.*'s proxy signature scheme as well as some other proxy schemes: the original signer can generate a valid proxy signature on any message of his own by forging another pair of proxy keys different from the pair that the proxy signer uses to sign. Consequently, many other proxy signatures [11, 12, 13] were shown to be also vulnerable to this kind of forgery attack.

In a summary, the current basic requirements having been set for a secure proxy signature are as follows,

- Verifiability: From a proxy signature, any verifier can be convinced of the original signers agreement on the signed message.
- Strong unforgeability: Only the designated proxy signer can create a valid proxy signature on behalf of the original signer. Anyone else, including the original signer, cannot do so.
- Strong identifiability: Anyone can determine the proxy signer's identity from a proxy signature.
- Strong undeniability: Once the proxy signer creates a valid proxy signature on behalf of the original signer, he cannot repudiate the signature creation against anyone else.
- Prevention of misuse: The proxy signer cannot use the proxy key for purposes other than generating a valid proxy signature. In case of misuse, the responsibility of the proxy signer should be determined explicitly.

Since proxy signatures and their derivatives can be used in many applications in reality, such as secure mobile agents [4, 5] and mobile communications [9], they have been receiving extensive research recently. In 2004, Xue *et al.* [14] proposed a threshold proxy signature (the XC scheme, for short), claiming that it could provide the properties of proxy protection, verifiability, strong identifiability, strong unforgeability, strong repudiability, distinguishability, known signers and prevention of misuse. Unfortunately, the XC scheme was shown by Guo *et al.* [3] to be vulnerable to an insider attack most recently. Also in 2004, Dai *et al.* [2] proposed a privacy-protecting proxy signature scheme where the message the original signer entrust to the proxy signer to sign on behalf of him is kept secret from the proxy signer during the generation of the proxy signature and could only be recovered by the receiver designated by the original signer. Therefore, the privacy of the original signer is protected. However, Cao *et al.* [1] pointed out that the receiver can cheat the proxy signer and obtain a proxy signature on any message in Dai *et al.*'s scheme and finally they presented an improvement (the CLX scheme, for short) to eliminate these weaknesses.

In this paper, we show that though the original signer cannot generate a valid proxy signature on any message of his choice in the XC scheme, he can individually produce another valid proxy signature on a message provided that he once gets a valid proxy signature on the message. The similar attack exists in the CLX scheme if the verifier does not check the originality of the proxy signer's proxy public key before verifying a proxy signature.

The rest of the paper is organised as follows. In the next section, some relevant notation are introduced. In Section 3 and 4, we show the attacks on the XC scheme and the CLX scheme, respectively. Some concluding remarks are made in Section 5.

2 Notation

The following notation will be used throughout this paper,

- p : a public large prime
- q : a public large prime factor of $p - 1$
- g : a public base element of order q in Z_p
- $h(\cdot)$: a public one-way hash function

3 Weakness in the XC Scheme

3.1 The XC Scheme [14]

The scheme consists of the following parties: a system authority (SA) whose task is to initialize the system, certification authority (CA) whose task is to generate the public key for each user, the original signer U_o , the proxy group of n proxy signers $G = \{U_1, U_2, \dots, U_n\}$, one designated clerk C whose task is to collect and then verify the individual proxy signatures generated by the proxy signers and finally construct the final threshold proxy signature, and a signature verifier. Let ID_i be the identifier of U_i . Assume that x_{CA} and y_{CA} are the private and public key of the CA, respectively, where $y_{CA} = g^{x_{CA}} \bmod p$. The warrant m_w records the proxy signers of the proxy group, parameters t and n , message type to sign, the valid delegation time, etc. *ASID* denotes the identities of the actual proxy signers.

The XC scheme consists of the following four phases:

Registration

Step 1. Each user U_i randomly selects an integer $t_i \in Z_q^*$, computes

$$v_i = g^{h(t_i, ID_i)} \bmod p,$$

and sends (v_i, ID_i) to the CA.

Step 2. Upon receiving (v_i, ID_i) , CA randomly selects $z_i \in Z_q^*$, computes

$$\begin{aligned} y_i &= v_i \cdot h(ID_i)^{-1} \cdot g^{z_i} \bmod p, \\ e_i &= z_i + h(y_i, ID_i) \cdot x_{CA} \bmod q, \end{aligned}$$

and returns (y_i, e_i) to U_i .

Step 3. U_i computes

$$x_i = e_i + h(t_i, ID_i) \bmod q,$$

and checks its validity by the following equation:

$$y_{CA}^{h(y_i, ID_i)} \cdot h(ID_i) \cdot y_i = g^{x_i} \bmod p.$$

Only if it holds does U_i accept (x_i, y_i) as his private and public key, respectively. Finally, CA publishes U_i 's public key y_i .

Proxy Share Generation

Step 1. The original signer chooses a random integer $k_0 \in Z_q^*$, computes

$$K_0 = g^{k_0} \bmod p, \tag{1}$$

$$\sigma_0 = k_0 \cdot K_0 + x_0 \cdot h(m_w) \bmod q, \tag{2}$$

and sends (m_w, K_0, σ_0) to each proxy signer.

Step 2. After receiving (m_w, K_0, σ_0) , each of proxy signers checks its validity by the equation below:

$$g^{\sigma_0} = K_0^{K_0} \cdot (y_{CA}^{h(y_o, ID_o)} \cdot h(ID_o) \cdot y_o)^{h(m_w)} \bmod p. \tag{3}$$

Only if it holds does each of proxy signers accept σ_0 as his proxy share.

Proxy Signature Issuing without Revealing Proxy Shares

The XC scheme allows any t or more proxy signers to represent the proxy group to cooperatively sign a message m on behalf of the original signer U_o . Let $G_p = \{U_{p_1}, U_{p_2}, \dots, U_{p_t}\}$ be the actual proxy signers for $t \leq t' \leq n$. To generate a threshold proxy signature, all members in G_p and the clerk C cooperatively perform the following steps:

Step 1. Each proxy signer $U_{p_i} \in G_p$ chooses a random integer $k_i \in Z_q^*$, computes

$$K_i = g^{k_i} \bmod p,$$

and sends it to the other $t' - 1$ proxy signers in G_p and the designated clerk C.

Step 2. Upon receiving K_j ($j = 1, 2, \dots, t', j \neq i$), each $U_{p_i} \in G_p$ computes:

$$K = \prod_{j=1}^t K_j \bmod p,$$

$$s_i = k_i \cdot K + (\sigma_0 \cdot t'^{-1} + x_{p_i}) \cdot h(m, ASID) \bmod q, \tag{4}$$

and sends s_i to C.

Step 3. For each received s_i ($i = 1, 2, \dots, t'$), C checks whether the following equation holds:

$$g^{s_i} = K_i^K \cdot \{[K_0^{K_0} \cdot (y_{CA}^{h(y_o, ID_o)} \cdot h(ID_o) \cdot y_o)^{h(m_w)}]^{t-1} \cdot y_{CA}^{h(y_i, ID_i)} \cdot h(ID_i) \times y_i\}^{h(m, ASID)} \bmod p. \tag{5}$$

Only if it holds does C accept that (K_i, s_i) is a valid individual proxy signature on m from U_{p_i} . If all the individual signatures are valid, then C computes $S = \sum_{j=1}^t s_j \bmod q$. Finally, the proxy signature on m is $(m_w, K_0, m, K, S, ASID)$.

Proxy Signature Verification

After receiving the proxy signature $(m_w, K_0, m, K, S, ASID)$, any verifier can verify its validity by the steps below:

Step 1. According to m_w and $ASID$, the verifier first obtain the value of t and n , the public keys of the original signer and proxy signers from CA and knows the number t' of the actual proxy signers. Then the verifier checks whether $t \leq t' \leq n$ holds. If it holds, the verifier continues the following steps. Otherwise, regard the threshold proxy signature invalid.

Step 2. The verifier checks its validity by checking

$$g^S = K^K \cdot [K_0^{K_0} \cdot y_{CA}^{(h(y_o, ID_o) + \sum_{j=1}^t h(y_j, ID_j))} \cdot h(ID_o)^{h(m_w)} \cdot y_o^{h(m_w)}] \times \prod_{j=1}^t (h(ID_j) \cdot y_j)^{h(m, ASID)} \bmod p. \quad (6)$$

The proxy signature is valid only if the Eqn.(6) holds.

3.2 Weakness

Strong unforgeability means that except the designated proxy signer, anyone else, including the original signer, cannot create a valid proxy signature on behalf of the original signer. Xue *et al.* claimed that the XC scheme meets the property of strong unforgeability. However, we note that it is not the case. Let's show the attack in details as follows.

In the phase of proxy share generation, note that the parameter k_0 in Eqn.(1) is chosen by the original signer and then K_0 is computed also by him. Suppose the original signer gets a valid proxy signature $(m_w, K_0, m, K, S, ASID)$ on the message m , then he chooses another random integer $a \in Z_q^*$, and computes

$$\begin{aligned} K'_0 &= K_0 \cdot g^a \bmod p, \\ S' &= S + [g^a \cdot K_0 \cdot (k_0 + a) - k_0 \cdot K_0] \cdot h(m, ASID) \bmod q, \end{aligned}$$

where k_0 is the randomly-chosen integer in Eqn.(1).

Now, $(m_w, K'_0, m, K, S', ASID)$ is another proxy signature on m , which is generated by the original signer, since

$$g^{S'} = K^K \cdot [K_0^{K_0} \cdot y_{CA}^{(h(y_o, ID_o) + \sum_{j=1}^t h(y_j, ID_j))} \cdot h(ID_o)^{h(m_w)} \cdot y_o^{h(m_w)}] \times \prod_{j=1}^t (h(ID_j) \cdot y_j)^{h(m, ASID)} \cdot g^{[g^a \cdot K_0 \cdot (k_0 + a) - k_0 \cdot K_0] \cdot h(m, ASID)} \bmod p$$

$$\begin{aligned}
 &= K^K \cdot [K_0^{K_0} \cdot g^{[g^a \cdot K_0 \cdot (k_0+a) - k_0 \cdot K_0]} \cdot y_{CA}^{(h(y_o, ID_o) + \sum_{j=1}^t h(y_j, ID_j))}] \times \\
 &\quad h(ID_o)^{h(m_w)} \cdot y_o^{h(m_w)} \cdot \prod_{j=1}^t (h(ID_j) \cdot y_j)^{h(m, ASID)} \pmod p \\
 &= K^K \cdot [K_0^{K_0} \cdot y_{CA}^{(h(y_o, ID_o) + \sum_{j=1}^t h(y_j, ID_j))}] \cdot h(ID_o)^{h(m_w)} \cdot y_o^{h(m_w)} \times \\
 &\quad \prod_{j=1}^t (h(ID_j) \cdot y_j)^{h(m, ASID)} \pmod p.
 \end{aligned}$$

Obviously, a verifier will mistakenly believe that the signature is really generated by the t impersonated proxy signers and the t impersonated proxy signers have no way to prove to a verifier that it is generated by the dishonest original signer. As a result, the original signer succeeds in forging another valid proxy signature on m under the name of his proxy signers.

Therefore, unlike what the proposers claimed, the XC scheme does not meet the property of strong unforgeability.

Remark: The above attack can be eliminated by hashing the K_0 with m_w as $h(m_w, K_0)$ and then replacing all the $h(m_w)$ in Eqns.(2),(3), (5) and (6) with $h(m_w, K_0)$. Furthermore, to prevent some potential attacks, we recommend that each $h(m, ASID)$ in Eqns.(4), (5) and (6) should be replaced by $h(m, K, ASID)$, where K is the one in Eqn.(4).

4 Weakness in the CLX Scheme

4.1 The CLX Scheme [1]

The scheme consists of the following parties: an original signer Alice, a proxy signer Bob and a receiver Cindy. Their secret and public key pairs are (x_A, y_A) , (x_B, y_B) and (x_C, y_C) , respectively, where $y_A = g^{x_A} \pmod p$, $y_B = g^{x_B} \pmod p$ and $y_C = g^{x_C} \pmod p$.

The CLX scheme consists of the following four phases:

Generation of Delegating Parameters

Step 1. The original signer Alice chooses a random integer $k \in Z_q^*$, and then computes

$$\begin{aligned}
 a &= g^k \pmod p, \\
 b &= m \cdot y_C^k \pmod p, \\
 \tilde{m} &= h(m), \\
 s_A &= x_A \cdot h(a, b, \tilde{m}, y_B, y_C) + k \pmod q.
 \end{aligned} \tag{7}$$

Thus, the receiver Cindy is designated by Alice through the form of receiver's public key y_C in the delegating parameters.

Step 2. Alice sends the delegating parameters $(a, b, \tilde{m}, y_A, y_B, y_C, s_A)$ to the proxy signer Bob.

Delivery and Verification of Delegating Parameters

The proxy signer Bob checks whether or not the equation $g^{s_A} = y_A^{h(a,b,\tilde{m},y_B,y_C)} \cdot a \bmod p$ holds. If it holds, Bob accepts it as a valid group of delegating parameters. Otherwise, reject it and request a valid one.

Signing by the Proxy Signer

The proxy signer Bob generates the proxy secret key x_P and corresponding public key as

$$\begin{aligned} x_P &= s_A + x_B \bmod q, \\ y_P &= g^{x_P} \bmod p = y_A^{h(a,b,\tilde{m},y_B,y_C)} \cdot a \cdot y_B \bmod p. \end{aligned}$$

Then, Bob can generate proxy signature on behalf of original signer by using x_P as follows: pick randomly an integer $\tilde{k} \in Z_q^*$, and computes

$$\begin{aligned} r &= \tilde{m} \cdot g^{-\tilde{k}} \bmod p, \\ \tilde{r} &= r \bmod q, \\ s &= \tilde{k} - \tilde{r} \cdot x_P \bmod q. \end{aligned}$$

Finally, Bob sends $(a, b, \tilde{m}, y_A, y_B, r, s)$ to Cindy.

Verification of the Proxy Signature

Firstly, by using her private key x_C , Cindy decrypts the message m from a and b as $m = \frac{b}{a^{x_C}} \bmod p$.

Then, Cindy verifies the proxy signature by checking if the following equation holds or not:

$$g^s \cdot y_P^{\tilde{r}} \cdot r = g^s \cdot (y_A^{h(a,b,\tilde{m},y_B,y_C)} \cdot a \cdot y_B)^{\tilde{r}} \cdot r = \tilde{m} \bmod p, \quad (8)$$

where $\tilde{r} = r \bmod q$ and $\tilde{m} = h(m)$.

If Eqn.(8) holds, Cindy accepts $(a, b, y_A, y_B, y_C, r, s)$ as a valid proxy signature of message m . Reject it, otherwise.

4.2 Weakness

In the phase of verification, Cindy must first verify the originality of Bob's proxy public key y_P and then verify the proxy signature according to the whole Eqn.(8), instead of just the later part (i.e., $g^s \cdot (y_A^{h(a,b,\tilde{m},y_B,y_C)} \cdot a \cdot y_B)^{\tilde{r}} \cdot r = \tilde{m} \bmod p$). Otherwise, after getting a valid proxy signature $(a, b, \tilde{m}, y_A, y_B, r, s)$ on the message m , the original signer can produce another valid proxy signature on m under the name of his proxy signer Bob as follows:

Step 1. The original signer Alice randomly chooses an integer $t \in Z_q^*$ and computes

$$\begin{aligned} \bar{b} &= b \cdot y_C^t \bmod p, \\ \bar{a} &= a \cdot g^t \bmod p, \\ \bar{s} &= s - \tilde{r} \cdot [t + x_A \cdot (h(\bar{a}, \bar{b}, \tilde{m}, y_B, y_C) - h(a, b, \tilde{m}, y_B, y_C))] \bmod q. \end{aligned}$$

Step 2. Alice sets the new proxy public key \bar{y}_P of his proxy signer Bob as

$$\bar{y}_P = y_A^{h(\bar{a}, \bar{b}, \tilde{m}, y_B, y_C)} \cdot \bar{a} \cdot y_B \text{ mod } p.$$

Now, $(\bar{a}, \bar{b}, \tilde{m}, y_A, y_B, r, \bar{s})$ is another proxy signature on m , which is generated by the original signer, for the two points below hold,

1. $\frac{\bar{b}}{\bar{a}^x} = \frac{b \cdot y_C^t}{(a \cdot g^t)^x} = m \text{ mod } p.$
2. Since the message m stays unchanged, its hash value \tilde{m} will not change, either. As a result, the following verifying equation holds:

$$\begin{aligned} &g^{\bar{s}} \cdot \bar{y}_P^{\tilde{r}} \cdot r \\ &= g^{\bar{s}} \cdot (y_A^{h(\bar{a}, \bar{b}, \tilde{m}, y_B, y_C)} \cdot \bar{a} \cdot y_B)^{\tilde{r}} \cdot r \text{ mod } p \\ &= g^s \cdot g^{-\tilde{r} \cdot [t + x_A \cdot (h(\bar{a}, \bar{b}, \tilde{m}, y_B, y_C) - h(a, b, \tilde{m}, y_B, y_C))]} \cdot (y_A^{h(\bar{a}, \bar{b}, \tilde{m}, y_B, y_C)} \cdot \bar{a} \cdot y_B)^{\tilde{r}} \cdot r \\ &= g^s \cdot (y_A^{h(a, b, \tilde{m}, y_B, y_C) - h(\bar{a}, \bar{b}, \tilde{m}, y_B, y_C)} \cdot y_A^{h(\bar{a}, \bar{b}, \tilde{m}, y_B, y_C)} \cdot \bar{a} \cdot g^{-t} \cdot y_B)^{\tilde{r}} \cdot r \\ &= g^s \cdot (y_A^{h(a, b, \tilde{m}, y_B, y_C)} \cdot a \cdot y_B)^{\tilde{r}} \cdot r \text{ mod } p \\ &= \tilde{m} \text{ mod } p, \end{aligned}$$

where $\tilde{r} = r \text{ mod } q.$

Therefore, $(\bar{a}, \bar{b}, \tilde{m}, y_A, y_B, r, \bar{s})$ will pass the verification if the verifier does not check the originality of Bob’s proxy public key or he verifies only according to the later part of Eqn.(8).

Please note the precondition that the designated verifier does not check the originality of Bob’s proxy public key y_P , otherwise, our attack will not succeed. However, it has made the CLX scheme vulnerable to potential attacks to some extent.

Remark: The above weakness can be eliminated by hashing a with m into \tilde{m} as $\tilde{m} = h(m, a)$ in Eqn.(7). And then, the new \tilde{m} is used throughout the rest of the scheme.

5 Concluding Remarks

In this paper, we show that the threshold proxy signature scheme [14] cannot prevent a dishonest original signer from producing individually another valid proxy signature on a message providing that he once gets a valid proxy signature on the message, and that a similar weakness will also arise in [1] if the designated verifier does not verify the originality of the proxy public key of the proxy signer before verifying a proxy signature. Therefore, both schemes are somewhat insecure.

Anyway, we should point out that the attacks may have only a little value in reality, for if the recipient sends the proxy signature generated by an original

signer to a proxy signer, the proxy signer can prove in some degree the dishonesty of the original signer by showing that he has two different groups of the proxy parameters that could only be generated by the original signer in the phase of proxy share generation. But in the point of academia, they significantly threaten the security of these two schemes.

Acknowledgments

The author is very grateful to the anonymous referees for their helpful suggestions to improve this work.

References

1. Cao. T., Lin. D., and Xue. R., Improved privacy-protecting proxy signature scheme, Proc. of AWCC'04 — Advanced Workshop on Content Computing, Chi-Hung Chi, and Kwok-Yan Lam (Eds.), Volume 3309 of Lecture Notes on Computer Science, pp.208–213, Springer-Verlag, 2004.
2. Dai. J., Yang. X., and Dong. J., A privacy-protecting proxy signature scheme and its application, Proc. of The 42nd annual Southeast regional conference, ACM Southeast Regional Conference, pp.203–206, 2004.
3. Guo. L., Wang. G., and Bao. F., On the security of a threshold proxy signature scheme using self-certified public keys, Proc. of CISC'05 — The SKLOIS conference on information security and cryptology, Higher Education Press of China. Dec. 15–17, 2005. Beijing, China. Archive available at <http://www.i2r.a-star.edu.sg/icsd/staff/guilin/publications.htm>
4. Lee. B., Kim. H., and Kim. K., Strong proxy signature and its applications, Proc. of SCIS'01 — 2001 Symposium on Cryptography and Information Security, pp. 603–608, Japan, 2001.
5. Lee. B., Kim. H., and Kim. K., Secure mobile agent using strong non-designated proxy signature, Proc. of ACISP'01 — 6th Australasian Conference on Information Security and Privacy, Vijay Varadharajan, and Yi Mu (Eds.), Volume 2119 of Lecture Notes on Computer Science, pp. 474–486. Springer-Verlag, 2001.
6. Li, L., Tzeng, S., Hwang, M.: Generalization of proxy signature-based on discrete logarithms, *Computers & Security*, Vol. 22(3),pp.245–255, Elsevier Science, 2003.
7. Mambo. M., Usuda. K., and Okamoto. E., Proxy signature: delegation of the power to sign messages, *IEICE Trans. Fundamentals*, Vol.E79-A: No.9, pp.1338–1353, 1996.
8. Mambo. M., Usuda. K., and Okamoto. E., Proxy signatures for delegating signing operation, Proc. of 3rd ACM Conference on Computer and Communications Security, pp.48–57, ACM press, 1996.
9. Park. H.U., and Lee. I.Y., A digital nominative proxy signature scheme for mobile communications, Proc. of ICICS'01 — Third International Conference on Information and Communications Security, S. Qing, T. Okamoto, and J. Zhou (Eds.), Volume 2229 of Lecture Notes on Computer Science, pp.451–455. Springer-Verlag, 2001.
10. Sun. H., and Hsieh. B., On the security of some proxy signature schemes, *Cryptology ePrint Archive: Report 2003/068*, available at <http://eprint.iacr.org/2003/068>

11. Tan. Z., Liu. Z., and Wang. M., On the security of some nonrepudiable threshold proxy signature schemes, Proc. of ISPEC'05 — First International Conference on Information Security Practice and Experience, Robert H. Deng, Feng Bao, Hwee-Hwa Pang, and Jianying Zhou (Eds.), Volume 3439 of Lecture Notes on Computer Science, pp.374–385. Springer-Verlag, 2005.
12. Wang. G., Bao. F., Zhou. J., and Deng. R.H., Security analysis of some proxy signatures, Proc. of ICISC'03 — 6th International Conference on Information Security and Cryptography, Jong In Lim, and Dong Hoon Lee (Eds.), Volume 2971 of Lecture Notes on Computer Science, pp.305–319, Springer-Verlag, 2003.
13. Wang. G., Bao. F., Zhou. J., and Deng. R.H., Comments on a Threshold Proxy Signature Scheme Based on the RSA Cryptosystem, Cryptology ePrint Archive: Report 2004/054, available at <http://eprint.iacr.org/2004/054>
14. Xue. Q., and Cao. Z., A threshold proxy signature scheme using self-certified public keys, Proc. of ISPA'04 — Second International Symposium on Parallel and Distributed Processing and Applications, Jiannong Cao, Laurence T. Yang, Minyi Guo, and Francis Lau (Eds.), Volume 3358 of Lecture Notes on Computer Science, pp.715–724, Springer-Verlag, 2004.
15. Zhang. K., Threshold proxy signature schemes, Proc. of ISW'97 — First International Workshop on Information Security, Eiji Okamoto, George I. Davida, and Masahiro Mambo (Eds.), Volume 1396 of Lecture Notes on Computer Science, pp.191–197, Springer-Verlag,1997.

A Proposal of Extension of FMS-Based Mechanism to Find Attack Paths

Byung-Ryong Kim¹ and Ki-Chang Kim²

¹ School of Computer Science and Engineering, Inha Univ., 253, YongHyun-Dong, Nam-Ku, Incheon, 402-751, Korea
doolyn@inha.ac.kr

² School of Information and Communication Engineering, Inha Univ., 253, YongHyun-Dong, Nam-Ku, Incheon, 402-751, Korea
kichang@inha.ac.kr

Abstract. With the increase of internet service providers(companies) for the rapidly growing numbers of internet users in recent years, malicious attackers has been growing too. Due to these attacks, corporate image can be impaired significantly by such damages as incredible service quality and unstable service, which can lead to fatal flaws. Among the malicious attacks, DoS(Denial-of-Service) is the most damaging and frequently reported form of internet attacks. Because DoS attacks employ IP spoofing to disguise the IP and hide the identity of the attacker's location, the correct address of attacker is not traceable only with the source IP address of packets received from damaged systems. Effective measures for the DoS attacks are not developed yet and even if defence is made for this attacks practically it is possible to repeatedly undergo attacks by the same attackers. In this point of view, in order to provide an effective countermeasure this study proposes mechanism to find out attack source by tracing the attack path using marking algorithms and then finding MAC address of attack source. In addition this study proposes technique to improve the packet arrival rate in marking algorithm and presents more effective measure with better performance to find attackers by enabling more prompt trace of the attack location

1 Introduction

Because of recent increase in internet users and internet service company's on-line commerce and corporate activities through the internet have been settled as one of routinely critical marketing means. Company activities through internet has been served as profit-generating business model with the internet publicity activities of companies including internet service providers and the development in electronic commerce. However the number of malicious attackers, who intend to prevent companies from performing on-line activities, has been increasing and it can cause immense damage to the target companies and service providers.

In an effort to detect location of attacker, IP traceback was approached through studies but because attacker spoofs its own IP there has been limit in tracing back IP. Attackers can hide its location spoofing the IP employing the point that unlike other attack forms DoS attack does not require creditable connection between the attack

system and the victim system and even packet having spoofed IP has enough effect on DoS attack[1,2,3,4].

Accordingly other than the technique to use source IP address recorded at packet, technique to detect attacker's location should be studied and marking type IP traceback technique was proposed as an alternative measure[5,6]. It detects the attack path by having every router on network mark its own IP address to packet passing through the router and using information on the victim system. This can detect to the first router where packet passes through out of the victim place but the disadvantage is that it cannot find the source location of the practical attack.

Therefore in the study we understand how to find attack path in current marking type IP traceback techniques and furthermore propose how to detect the true attack source with MAC address by marking MAC address of attack source.

For this each router marks the MAC address of front end together with its own IP address and it is proposed as follows; mechanism to perform integrity test on MAC address just as the integrity test performed on IP address in the victim system, and minimize the number of packets needed to find out attack path, which was revealed as weakness.

This paper is composed of the following parts. In chapter 2, concerned research methods are introduced. Dos attack, attacker traceback, and marking-base IP traceback are introduced, and problems in current marking algorithm-base IP traceback techniques are presented. In chapter 3 in order to solve one of problems of traceback in detecting attack source we propose the solution using MAC address and technique to improve packet arrival rate. In chapter 4, Technique to Create MAC-Slice Combination and Check Integrity are shown and chapter 5 Correction of Sampling Probability p . Finally, chapter 6, 7 conclusion and performance are described.

2 Consideration of Existing Studies

Current security systems have passively coped with attacks by blocking the attacks and decreasing system damage. But more active mechanism to trace the source of attack by tracing back the attacker becomes influential as new security mechanism. So mechanisms to trace back attackers are discussed in diverse perspectives. There have been many presented mechanisms to trace back marks of attackers including analyzing systems log[9,10], logging[8], ingress[7], filtering, link testing[12], ICMP traceback[11] and so forth. Each has its strong and weak points and it was not sufficient to trace back attackers.

In this respect marking algorithm is newly proposed. It is possible to trace back attack path by having intermediate router mark its IP address to packet.

In this mechanism, referred to as Fragment Marking Scheme(FMS), router marks that packet appointed as given probability, p or less, not all the packet, passes through the router. Namely it marks the router IP to source field for additional data to show path of packet which is selected with probability of p or less at each router and distance information, the number of hops is set to 0. When the probability to select the packet at the next router through which the packet passes is p or more, the last router marked at the packet is the IP of the router. The record on both the source and last IP undergoes the following; when passing through the next router, if probability is p or

more, the number of hops is added by the number of routers passed through from the source router and if less than p , this record is disregarded and source is recorded again. This series of processes are reliable to find out the source of the real packet although modification is made to the source address where packet was generated on purpose by increasing the random rate. This mechanism has come from the idea that the part to record the data is the identification field marking the packet's identity for the separation of packet's IP header and the rate to use the part is statistically 0.25%. Fig.1 is newly defined identification field.

Offset information	Distance information	Fragment information
0 2 3	7 8	15

Fig. 1. Identification field configuration of redefined IP header

In current marking algorithm, router data is sent to identification field of packet's IP header, router data is sent in slice and marking process at router is processed as probability by sampling basis. In order for router to mark its IP address to packet, identification field(16 bit) of IP header is used.

Using the router's IP address, R and bit algorithm to the IP address 64 bit R is created by bitinterleaving the hash(R) value and o (displacement data, the random number)th slice is loaded to packet after the value is divided into 8 slices. So distance data, the slice of IP address, and the displacement data indicating the slice's location are marked. The packets marking the IP address of passed router are composed by distance data at the victim system and if the IP address obtained from the combination is found to be right router's IP address, it will be recorded to path tree. So attack path can be the path tree composed through the above described processes. Fig 2 shows the real attack path which can be found through the existing marking algorithm among many transmission paths of packet.

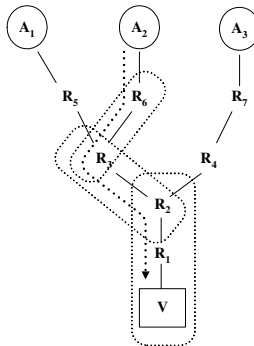


Fig. 2. Packet's many transmission and attack paths

As shown on Fig 2 with current marking algorithm to the first router of attack path can be found. Although attack paths are found and using the paths measures can be devised to cope with DoS attack, source of attack(A_2) cannot be found. To the first router through which packet passed it is possible to trace but there is no way to trace attack source here. Also due to the algorithm's characteristic that the first data is lost(marking the router of new router) in the middle by each router till packet arrives victim systems, it is relatively low for packet having the first router data to arrive to the victim system. The more the number of hops from attacker to the victim system, the smaller the arrival rate of the packet. When the probability value of each router is p and the number of hops from router to the victim system is d , the probability for packet data to arrive is written in equation(1);

$$E(x) = \frac{1}{p(1-p)^{d-1}} \quad (1)$$

3 Mechanism to Detect Attack Source

Router's location is marked by marking the router's IP address to packet in existing marking algorithms, but because only the router data on path(IP address) is included in packet, it is traceable only to the first router on path. Hence it is not possible to trace the location of attack source in the existing mechanisms.

In this paper we intend to mark not only router's IP address but also the MAC address of front end(the previous router or attack source). While there is advantage in marking MAC address there also is problem. The problem is that it can send 32 bit IP address and 48 bit MAC address in same way. In case of IP address, making 32 bit hash value, 64 bit is made adding the hash value to the existing IP address, transmission is performed in 8 slices and integrity can be checked in reverse process. But in case that transmission is done in 8 slices after making 48 bit MAC address into 64 bit, each router should process two stages of marking its IP address and the front end's MAC address and one router leads to consuming two distance data to discern each stage. Because of this the number of hops to be traced back is limited to 16 which is the half in the existing mechanism and which is a great obstacle to the original aim to trace attack source. In this way new mechanism is required to allow to mark MAC address and at the same time not to decrease the traceback scope under current mechanism.

To load router's IP and MAC address data together to identification field as in marking algorithm, check the integrity on the two data in the victim system, and maintain the number of hops of attack path that was possible to trace in current mechanism, this paper proposes that IP and MAC addresses are transmitted in slices after making the addresses into combination of 56 bit slices. Here mechanism to check the integrity should be added in making 56 bit combination. For this identification field of IP header is to be recombined. Proposed recombination technique is to combine 16 bit of identification field into 3 bit displacement data field, 6 bit distance data field, and 7 bit slice data field. This can be illustrated as Fig 3.

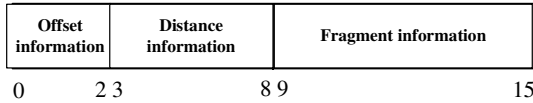


Fig. 3. Identification field configuration of proposed IP header

Mechanism to create IP slice and MAC slice to be loaded to slice data field of IP header identification field and check the integrity is proposed as follows. Unlike in current marking algorithms that 64 bit IP-slice combination is made by adding hash 32 bit to 32 bit IP address. in this paper 56 bit IP-slice combination is composed and technique to check the integrity of IP address is also to be included. Lemmas 1 to show process to create IP-slice combination

Lemmas 1. IP-slice combination creation algorithm

Input : 32 bit IP address

Output: 56 bit IP-slice combination

- Divide 32 bit IP address(R) into 4 parts and call it A, B, C, D each.
- Perform XOR (\oplus) over the former two 8 bits(A, B) and the latter two 8 bits(C, D) and make two new 8 bits(E, F) and then perform XOR to the two(E, F), create one new 8 bit and the value is called G .

$$E = A \oplus B$$

$$F = C \oplus D$$

$$G = E \oplus F$$

24 bit(H) is created by combining newly created three 8 bits.

$$H = E + F + G$$

- 56 bit is created adding 32 bit R and the new 24 bit H .(Separating 4 bit at R and 3 bit at H from the front, adding the two makes 7 bit. Repeat the process 8 times then 56 bit R' is created.)

Here the process to create the 56 bit R' is defined as BitInterleave1.

Divide the created R' into 8 slices and load them to packet according to probability.

In the victim system a complete IP address is combined after recombination over 8 packets having the same value of distance data, and here it should be checked whether the combined IP address is correct or incorrect mixed with other slice in the process.

Lemmas 2, the checking stage, is the reverse process of the above proposed Lemmas 1.

Lemmas 2. Integrity Check Algorithm of IP address

Input: 56 bit IP-slice combination

Output: 32 bit IP address

- BitDeinterleave and combine the 5, 6, 7 multiple bit of 56 bit's IP-slice combination(R') and make 24 bit and this is called H .
- Divide this 24 bit by 3 and call E, F, G each and where $G' = E \oplus F$, perform process 3.

Where $G \neq E \oplus F$, delete R' and re-preform with new slice combination from process 1.

- Combine created three 8 bits and call it H .

$$H = E + F + G$$

- Here 32 bit excluding R' at H becomes IP address.
- Divide R' by 4 and perform XOR over the former two 8 bits(A, B) and the later two 8 bits (C, D) each. After making two new 8 bits(E', F') perform XOR over these two 8 bits, create one new 8 bit and call the value G' .

$$E' = A \oplus B$$

$$F' = C \oplus D$$

$$G' = E' \oplus F'$$

- After combining three newly created 8 bits create 24 bit(H').

$$H' = E' + F' + G'$$

Where $H \neq H'$, accept the R' is correctly transmitted combined value in correct order, and where $H \neq H'$, delete R' and perform from process 1 with new combination slice.

4 Technique to Create MAC-Slice Combination and Check Integrity

As each router on attack path marks its own IP address to packet, if it marks the MAC address of front end also, it needs to check the integrity of the transmitted MAC address as the integrity of IP address transmitted to packet is checked in the victim system. In addition by having the intermediate router mark the MAC address of the front end, practically each router consumes two distance data(its own IP address - distance data 0, MAC address of the front end - distance data 1). This causes a new problem of decreasing the number of traceable hops under the current mechanism to the half.

To fill out this problem, in this study adjusting the configuration of identification field and making distance data field 6 bit, the number of traceable hops can be maintained as in current marking algorithm by composing slice data field with 7 bit. Also the problem of decreasing slice data field by 1 bit was settled by proposing a technique to check integrity after adjusting the existing 64 bit slice combination into 56 bit.

Lemmas 3, an algorithm to show the process to create MAC-Slice combination, explains the process with the example.

Lemmas 3. MAC-slice combination creation

Input: 48 bit MAC address

Output: 56 bit MAC-slice combination

- Divide 48 bit MAC address (M) by 6, make it 6 parts, and call it A, B, C, D, E, F each.
- XOR each 8 bit, create new 8 bit and call this value G .

$$G = A \oplus B \oplus C \oplus D \oplus E \oplus F$$

- BitInterleave each bit of new 8 bit, G , to the 7th multiple place of MAC address, create 56 bit, M . Here define the process of creating M as BitInterleave2.

Divide the created M by 8 slices, and load them to packet according to probability.

5 The Correction of Sampling Probability p

Sampling probability p is the standard to determine whether to neglect the existing value for packet obtained at each router and mark the new value to identification field with its IP address or to XOR and mark IP address to existing value.

In current marking algorithm probability p was fixed at each router. By algorithm's nature, if packet received by router is smaller than p with standard of probability p , the previous data is lost and current router's data is newly marked. On average if it is set to $p=0.5$, whenever passing through router, the number of packets having the first data is decreased by the half.

Finally to configure attack path under the victim system the smallest number of packets among the number of packets required to construct attack path contains the first router on the attack path. This is because the previous data is lost whenever packet passes through router so the longer router from the victim system is, the harder it gets to send the data to the victim system. It is absolutely due to the probability p set to each router. Since the probability value is fixed packet's previous data gets lost as it passes by hop at regular rate.

It was found that packet arrival rate can be improved if probability rate is corrected to give weight to long-distance packet since probability value of every router is regularly set so that the longer the distance of packet is the lower the arrival rate to the victim system. For this we propose the following mechanism to give weight to probability by packet's distance data.

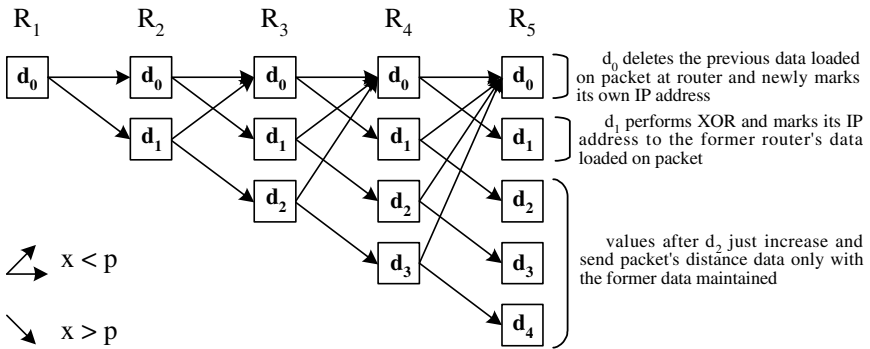


Fig. 4. Packet flow and data in marking algorithm

Fig 4 shows that router data saved at packet is newly created or changed in the course that packet starts sending after router's marking its data to the packet at router R_1 and arrives to Router R_5 thru each router and also shows each packet discerned by router data saved at such packet.

In Fig 4 d_0 deletes the previous data loaded on packet at router and newly marks its own IP address, d_1 performs XOR and marks its IP address to the former router's data loaded on packet, and values after d_2 just increase and send packet's distance data only with the former data maintained.

In Fig 4 arrow means to send to the next router applying probability p , for example if probability p applies to d_0 of R_1 , the value is d_0 and d_1 of R_2 . Here if applying probability produces random number and $x < p$, follow process(1) in Fig 4, if packet distance data is 0 and $x > p$, follow process(2), and if greater than 0 and $x > p$, follow process(3)

It can be described in the following equation.

$$R_2[d_0] = R_1[d_0] \times p \tag{2}$$

$$R_2[d_1] = R_1[d_0] \times (1 - p) \tag{3}$$

In current mechanism, probability p is fixed at every router. Now by giving weight to the probability by distance, data loss and arrival rate is to be improved. It is help-less to lose data owned whenever passing through router but it is necessary to minimize time to spend in finding attack path by having packet with distant data arrive at the victim system as much as possible.

Value to indicate the number of packets arriving at each router, X , is defined below.

Definition. Where number of hops is n , the number of packets arriving at router with packet, whose distance is i , is expressed in X_i^{n+1} .

Also function j to probability p is defined in the following equation to weigh by distance data.

$$f_i = \frac{1}{(i+1) \times 2} \quad (i=\text{distance}) \tag{4}$$

With the two above definitions value to hop can be expressed in the equation below.

Where $n=0$,

$$X_0^1 = X_0^0 \cdot f_0$$

$$X_1^1 = X_0^0 \cdot (1 - f_0)$$

Where $n=1$,

$$X_0^2 = X_0^1 \cdot f_0 + X_1^1 \cdot f_1$$

$$X_1^2 = X_0^1 \cdot (1 - f_0)$$

$$X_2^2 = X_1^1 \cdot (1 - f_1)$$

⋮
⋮
⋮

Combining above equations it can be expressed in the following function equations.

$$X_0^{n+1} = X_0^n \cdot f_0 + X_1^n \cdot f_1 \dots X_n^n \cdot f_n \tag{5}$$

$$X_i^{n+1} = X_{i-1}^n \cdot f_{i-1} \text{ for } i=1, \dots, n+1 \tag{6}$$

Without setting the probability fixed as expressed above, if weighing by packet's distance data, whenever passing through router the packet's arrival rate with the previous data, which was greatly diminished, will be largely improved.

6 Performance

Mechanism to weigh router's probability p by packet's distance data was proposed above. To confirm whether equation 4 is optimized or not, function equation to probability p is changed to the following function having two coefficients a, b .

$$f_i = \frac{1}{ai+b} \text{ (} i=\text{distance data, } a \geq 0, b > 0 \text{)} \tag{7}$$

First it will be found out how packet's arrival rate is changed as hop gets longer. Fig 5 shows the packet's arrival rate changed by hop and distance data when probability is fixed to $p=0.5$ at each router.

distance \ hops		hops					
		1	2	3	4	5	6
packet's arrival rate	d=0	50	50	50	50	50	50
	d=1	50	25	25	25	25	25
	d=2		25	12.5	12.5	12.5	12.5
	d=3			12.5	6.25	6.25	6.25
	d=4				6.25	3.125	3.125
	d=5					3.125	1.5625
	d=6						1.5625

Fig. 5. Packet arrival rate by hop and distance data(p fixed) (Unit %)

As shown on Fig 5 as hop increases the arrival rate of packet arriving to remote router diminishes to half each. With some distance data, packet's arrival rate(R_i) is changed by the following equation.

$$R_i = \frac{100}{2^i} \text{ (} i=\text{distance data)} \tag{8}$$

To sum up this, it can be found the more hops increase and the bigger distance data gets, the dramatically smaller packet arrival rate. The packet arrival rate by hop and distance data obtained from the application of equation 4, function proposed to probability p , in the results packet arrival rate has been improved by flexibly weighing p by distance data rather than by fixing the probability p .

7 Conclusion

There have been significant problems in tracing back attack path with marking algorithm among IP traceback mechanisms stressed as countermeasure against DoS attack

which has caused great damage to internet service companies. It was effectively improved to solve problems in detecting attack source and requiring great deal of packets to find attack path to cope with DoS attack by proposing mechanism to find MAC address of attack source and improve packet arrival rate by weighing router's sampling probability value. While the attack source could not be found in current marking algorithm, in the proposed mechanism MAC address of attack source can be found by marking MAC address at intermediate router and tracing back attack path in the victim system and real packet arrival rate and the deviation of the rate can be greatly reduced by changing the value of router's sampling probability from fixed value into flexible value to give weight by distance data of packet.

Acknowledgements

This work was supported by INHA UNIVERSITY Research Grant.

References

1. Computer Emergency Response Team (CERT), CERT Advisory CA-1995-01 IP Spoofing Attacks and Hijacked Terminal Connections, <http://www.cert.org/advisories/CA-1995-01.html>, Jan. 1995.
2. Computer Emergency Response Team (CERT), CERT Advisory CA-2000-01 Denial-of-service developments, <http://www.cert.org/advisories/CA-2000-01.html>, Jan. 2000.
3. S. A. Crosby, D. S. Wallach, Denial of Service via Algorithmic Complexity Attacks, in: Proceedings of the 12th USENIX Security Symposium, 2003.
4. Project IDS - Intrusion Detection System, <http://www.cs.columbia.edu/ids/index.html>, 2002.
5. Dawn Xiaodong Song and Adrian Perrig, Advanced and Authenticated Marking Schemes for IP Traceback, in Proc. IEEE INFOCOM, April. 2001.
6. Stefan Savage, David Wetherall, Anna Karlin, and Tom Anderson, Practical network support for IP traceback, in Proc. of ACM SIGCOMM, pp. 295-306, Aug. 2000.
7. P. Ferguson and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks Which Employ IP Source Address Spoofing, RFC 2267, Jan. 1998.
8. G. Sager. Security Fun with Ocxmon and Cflowd. Presentation at the Internet 2 Working Group, Nov. 1998.
9. Computer Emergency Response Team (CERT), <http://www.cert.org/index.html>, 2002.
10. David A. Curry, UNIX System Security," Addison Wesley, pp.36-80, 1992.
11. S. M. Dellovin, The ICMP Traceback Messages, Internet Draft: draft-bellovin-itrace-00.txt, <http://www.research.att.com/~smb>, Mar. 2000.
12. R. Stone, CenterTrack: An IP Overlay Network for Tracking DoS Floods, In to appear in Proceedings of the 2000 USENIX Security Symposium, Denver, CO, July. 2000.

Comparative Analysis of IPv6 VPN Transition in NEMO Environments*

Hyung-Jin Lim, Dong-Young Lee, and Tai-Myoung Chung

Internet Management Technology Laboratory and
Cemi: Center for Emergency Medical Informatics,
School of Information and Communication Engineering,
Sungkyunkwan University,
Chunchun-dong 300, Jangan-gu, Suwon, Kyunggi-do,
Republic of Korea
{hjlim, dylee, tmchung}@rtlab.skku.ac.kr

Abstract. IPv6 deployments use different NGtrans mechanisms, depending on the network situation. Therefore, a mobile network should be able to initiate secure connectivity with the corresponding home network. The correlation between transition mechanisms used in existing network and transition mechanisms supported by the mobile network, is an important factor in secure connectivity with a corresponding home network. In this paper, VPN scenarios applicable to mobile networks during the transition to IPv6, have been analyzed. In addition, performance costs have also been evaluated, in order to establish connectivity according to VPN models with each transition phase. In addition, this paper analyzes factors affecting performance, through numeric analysis for NEMO VPN model, under an IPv6 transition environment. As shown in our study, a NEMO VPN creating hierarchical or sequential tunnels should be proposed after careful evaluation of not only security vulnerabilities but also performance requirements.

1 Introduction

Present Internet deployment consists of an IPv4-based network. However, this network environment cannot cope with demands for various services and expanding users resulting both from network technological development and the rapid spread of the Internet. To overcome this demand, the IETF has proposed the IPv6 protocol, IPv6 includes several excellent features such as newly recognized routing efficiency, mobility support, QoS, auto-configuration. Additionally, the IPv6 protocol solves the exhaustion of IPv4 addresses. In IPv4, IPSec is implemented using an additional module, to process the AH and ESP header. However, in IPv6, IPSec is an integrated component, as regular use is expected. A considerably long period of time will elapse before IPv6 in networks is developed, during this time it will coexist with the present IPv4 protocol.

* This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment).

For natural transition to a full IPv6 network, the IETF NGtrans working group has developed a variety of transition mechanisms [1]. In contrast to situations with pure IPv4 or IPv6 networks, the availability of IPsec for VPN composition in an IPv6 network development process is an important factor in choosing a transition mechanism.

The IETF NEMO working group has been doing research relating to mobility, based on a mobile router, as representative of an IPv6 application, mainly focusing on improvements of basic protocol, route optimization (RO), and multi-homing [2-4]. Since NEMO supports basic mobility in the network unit, an extended application of existing VPN models may also be required. This means that analysis of such VPN models in the transition phase to IPv6 is required, even in the NEMO research domain. Therefore, in this paper, VPN scenarios applicable under mobile network environments are evaluated. In Section 2, the IPv6 transition phase and scenarios are analyzed, to establish the VPN under the NEMO environment. Section 3 and Section 4 covers the evaluation and analysis of costs expected in accordance with each scenario.

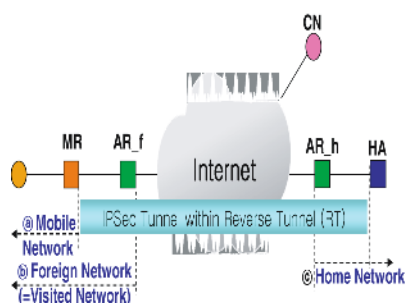


Fig. 1. Mobile network topology and connectivity

①	②	③	Internet Connectivity	MR Requirement
①	IPV4	IPV4	v6: 6 to 6 over 4 v4: Net-T, 4 to 4 over 6	6 to 6 over 4, Must be configure from IPv4 Router.
②	Dual Stack	IPV4	v6: 6 to 6 over 4, Net-T, v4: 4 to 4 over 6, Host-T.	Transition mechanism with AR f can be used.
③	IPV6		v6: 6 to 6 over 4, Net-T, v4: Host-T.	
④	IPV4	Dual	the same with ①	the same with ①
⑤	IPV4	IPV6	the same with ①	

* Net-T: Network translation, (i.e., NAT-PT), Host-T: Host translation (i.e., BIA, BJS),
VN: Visited Network

2 IPv6 VPN Transition Scenario in NEMO Environment

The IETF NEMO proposes the IPv6 extension protocol as supporting mobility for MR. Fig. 1 presents the NEMO basic topology. Under the IPv6 transition environment, IP versions of the visited network and Internet determine the MR requirements in establishing connectivity. That is, a total of 27 combinations are made possible due to IP versions by MR, visited networks, and the Internet.

It is assumed that, in the case of a visited network supporting only IPv4, the MR is given CoA from AR_f and is a mechanism for setting up IPv6 CoA [5]. It is also assumed that such a visited network can support NEMO, irrespective of its IP version. Additionally, the MR should to recognize the version of the visited network.

2.1 Connectivity

Fig. 1 demonstrates a basic NEMO VPN model, in which reverse tunneling connects a MR with HA. An IPsec tunnel exists within the reverse tunnel, therefore, the packets contain an external IP encapsulation header and IPsec extension header. In this study, the main transition phases of the five cases of NEMO are considered in the

IPv6 transition environment. The cases where the IPv6 transition mechanism by MR is not required, enabling IPv6 communications is excluded. It is possible that the Internet and a visited network (AR_f) support various IP versions as only-IPv6, only-IPv4 and Dual-stack. A transition mechanism required for MR and AR to set up connectivity according to their network status is presented in Fig 1. Connectivity of the Internet refers to a transition mechanism where a host in a foreign network may be used to form connectivity with an arbitrary host within a network connected to the Internet.

If a visited network (AR_f) supports dual stack, its inner network device or end-hosts are able to enable IPv6 or IPv4. However, if the visited network (AR_f) is an only-IPv4 node, a transition mechanism as Bump In the API (BIA), Bump In the Stack (BIS) and Network Address Translation-Protocol Translation (NAT-PT) is required for its inner IPv6 devices to communicate with other IPv6 devices. It is assumed that AR_h is a network device in support of IPv6, as is the case with MRs. Different IP version of the Internet require different transition mechanism for MR or a visited network to create connectivity. Under this mobile environment, when the mobile network has been involved in a foreign network, it may demand establishment of a different connectivity, in accordance with transition mechanism the foreign network suggests.

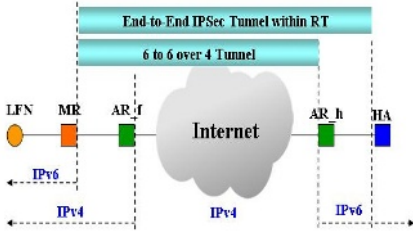


Fig. 2. Configuration of VPN in Fig. 1_table-①

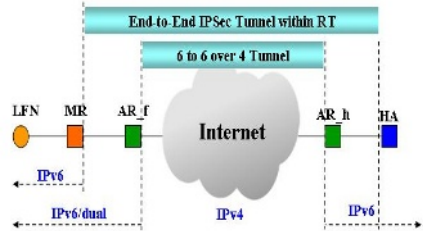


Fig. 3. Configuration of VPN in Fig. 1_table-② & ③

2.2 NEMO VPN in IPv6 Transition

An additional IPSec tunnel should be created to guarantee secure connectivity, irrespective of whether a transition mechanism exists or is used under such an environment. From the perspective of a VPN, a remote access VPN model can be applied to a mobile environment. However, a Gateway-to-Gateway VPN model (=intranet VPN) is suitable in a MR, since it is a network. Also, we took into account only the VPN connectivity based on not host-based but network-based translation mechanisms since IPv6-based MRs include a method to set up connectivity with its IPv6-based home network. The home network should establish connectivity with its own IPv6-based MR. Therefore, the case where an IPv6 host within a MR set up connectivity with an IPv4 end-host within a network can also be considered. In this case, the MR or the IPv6 host may require a translation mechanism.

Fig. 2 demonstrates that when an IPSec tunnel for end-to-end secure channel is formed, the MR sets up connectivity with its own home network via an IPv4 carrier, through a tunneling mechanism supporting 6-to-6 over 4 connectivity. In Fig. 3, IPv6 is supported in a visited network. From a MR perspective, IPv6-based NEMO is made possible, without any transition mechanism. However, a 6-to-6 over 4 tunneling mechanism is required because the Internet still consists of an IPv4 carrier. In contrast to the case in Fig. 2, since the visited network has already used such a tunneling mechanism, a tunneling mechanism supported by its AR_f can be used. This implies that the MR does not require provision of a separate mechanism.

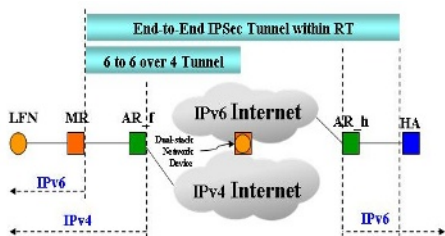


Fig. 4. The case of Fig. 1_table 1-④

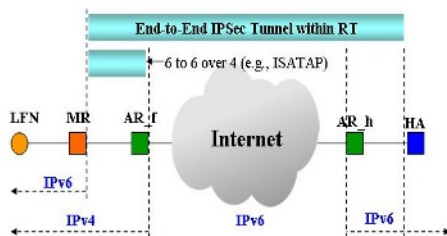


Fig. 5. The case of Fig. 1_table 1-⑤

Fig. 4 demonstrates that two IP versions on the ISP carrier both exist, on whose contact boundary a dual stack network device or a network-based translator can be located. For connectivity required from a NEMO perspective, the MR requires the 6-to-6 over 4 tunneling mechanism over the IPv4 carrier to the dual stack network device. Fig. 5 refers to a case where the Internet infrastructure offers IPv6, in which case the Internet is dedicated to IPv6 or supports a dual stack network device. In that case, since only a visited network uses IPv4 version, connectivity with the MR's home network is possible if a tunneling mechanism is used, such as ISATAP.

3 Evaluation of NEMO VPN with NGtrans Mechanism

3.1 Consideration and NEMO VPN Model

In this section, an end-to-end VPN model and a transition mechanism are analyzed in a NEMO environment, from the view of processing cost. Additionally, this section also concerns costs and influences triggered by creating a VPN in the 5 models in the preceding section and IPv6-based NEMO VPN without the transition mechanism.

When VPN tunneling is established, the cost of processing according to the IPSec mode and service are included. Likewise, when the NEMO VPN is included in IPv6 transition environments, the performance in data transmission includes the packet translation cost between different IP versions, cost of IP tunneling, and cost of VPN tunneling related to the cryptography process. The packet translation cost between different IP versions refers to the processing cost of translation mechanisms, and the cost of IP tunneling refers to the processing cost of tunneling mechanisms, composed of the various connectivity and tunnel type (i.e., 6-to-6 over 4 or 4-to-4 over 6).

In this paper, the performance parameters considering the transmission efficiency of the VPN are extracted from previous experimental results [6-11] and the cost effects of each VPN model are evaluated in NGtrans environments. The cost estimation of the VPN constitution is conducted by the per packet cryptography process. In order to evaluate the processing overhead in applying VPN in a transition environment, the following parameters in Table 1, were considered.

Table 1. VPN’s Performance Parameter in NGtrans Environment

Parameter	Description	Value	
C_{Sh}	Processing cost for Security header	1.5~7.5	
C_{IP}	Processing cost for IP encapsulation header	1.6	
v_{con}	v_{con} means weight value as increase ratio for IPv4 transfer packet amount by the result of TCP payload reduction,	V_4	1
		V_6	1.0137
		V_{4sec}	1.015~0.0164
		V_{6sec}	1.0287~0.0301
		V_{4IP}	1.0137
		V_{6IP}	1.041
bw_a	Bandwidth of the access network in local area network	100Mbps	
bw_i	Bandwidth of the backbone network	E1(1.544Mbps)	
fg	Weighted value by a fragmentation on network device	0.1(default)	
p_{kt}	Transfer packet’s amount per host	64k~1Mbyte	
l_p	Latency include processing of network and data link layer	500 μ SEC	
N	Hop count in public network	20~30hop	
n	Hop count within local network	1~3hop	
N_n	End-host’s number with VPN flow	1~100flow	

In the VPN transmission mode, only the transport mode is evaluated. When evaluating the tunnel mode, the expansion of transmission packets by adding the external header, fragmentation and additional cost of cryptography should be considered. In addition, connectivity only using basic transition mechanism operation, in conjunction with NEMO VPN models, is considered when estimating the cost. It is assumed that the configurations for security association and IP tunneling are already established for initiating communication.

3.2 Evaluation of NEMO VPN Model

Fig. 1 presents a basic model of NEMO under a pure IPv6 environment, where all network devices and end-hosts only support IPv6. It is assumed that a Gateway-to-Gateway VPN model is applied between the MR and its home network, because of the assumption that the inner MR is a secure section. Additionally, although the IPSec transport mode from the MR end-host can be recommended, the

transport mode is considered in the only MR, in order to analyze relative VPN performance influence implemented in each model. Both transport mode processing in the end-host, and tunnel mode processing in the MR are used. Therefore, VPN processing costs incurred for transmission packets are evaluated from the end-host in the MR to the end-host located in the MR's home network as follows:

$$V_{basic} = (2P_{rate}(C_{sh}(v_{6sec} - 1) + C_{ip}v_{6sec}(v_{6ip} - 1)(1 + fg)) + (v_{6ip}v_{6sec}(t_i + 2t_1)))N_n$$

Reverse tunnel and IPSec tunnel processing is performed on MR and HA. Processing costs required for tunneling are calculated in accordance with $P_{rate}(C_{sh}(v_{6sec} - 1) + C_{ip}v_{6sec}(v_{6ip} - 1)(1 + fg))$. Packets transmitted from an end-host first process the IPSec header in the MR, and then establish the reverse tunnel, in which case additional processing costs are introduced as much as $P_{rate}C_{sh}(v_{6sec} - 1)$ and $P_{rate}C_{ip}v_{6sec}(v_{6ip} - 1)$. Since MR and HA perform the functions of a router, fragmentation constant $(1 + fg)$ is additionally considered along with established reverse tunneling.

In this case, P_{rate} refers to the number of cryptographic processes for the amount of transmission packets on the basis of an Ethernet payload of 1500bytes, which means $P_{rate} = P_{kt} / MTU$ [12]. In IPSec, the frequency packet processing may vary depending upon the algorithm (e.g., DES, SHA1, MD5 etc) that is applied to AH and ESP. However, it is assumed that the number per MTU is identical. In particular, additional headers are created by v_{6sec} than in the case of the Ethernet MTU basis, as can be compared with that of IPv4. Therefore, local hosts should perform additional processing as $P_{rate}C_{sh}(v_{6sec} - 1)$. Encrypted packets travel through n local network hops and N Internet carrier hops, and finally to the VPN gateway, in which case each hop requires transmission processing costs as much as $t_i (= (l_p + (P_{kt} / bw_i))N)$ and $t_1 (= (l_p + (P_{kt} / bw_i)))$, respectively. At this time, weight is placed upon the amount of an increase in packets, in accordance with the version of connectivity and type of tunneling (i.e., VPN tunnel or IP-in-IP tunnel). This means that the local network of the remote host, the Internet carrier, and the home network VPN gateway, incur transmission costs as much as $v_{6ip}v_{6sec}$.

Finally, in accordance with an increase in the VPN end-host, Nn is taken into account. It is assumed that each end-host only has a single flow and has the same increase rate in its packets. Therefore, there is no influence upon the processing costs of individual hosts even if the number of end-hosts increases. In Fig. 2, aside from 6-to-6over4 tunneling between MR and AR_h to set up connectivity, a reverse tunnel including an end-to-end VPN tunnel from MR to HA is established, in which case the costs incurred are as follows:

$$V_{ml} = (P_{rate}(C_{sh}(v_{6sec} - 1)(2 + fg) + 2P_{rate}C_{ip}v_{6sec}(v_{6ip} - 1) + 2P_{rate}C_{ip}v_{6sec}v_{6ip}(v_{4ip} - 1)(1 + fg) + v_{6sec}v_{6ip}v_{4ip}(t_i + t_1) + v_{6sec}v_{6ip}t_1)N_n$$

Where MR requires a cost as much as $P_{rate}C_{sh}(v_{6sec}-1)+P_{rate}C_{ip}v_{6sec}(v_{6ip}-1)$ to process the reverse tunnel and the VPN tunnel. Next, the 6-to-6 over 4 tunnel processes $P_{rate}C_{ip}v_{6sec}v_{6ip}(v_{4ip}-1)(1+fg)$ by the level of increased packet overhead. These packets cross the Internet to reach the access router AR_h, in which each hop incurs costs as much as $v_{6sec}v_{6ip}v_{4ip}(t_i+t_1)$. AR_h also incurs processing costs as much as $P_{rate}C_{ip}v_{6sec}v_{6ip}(v_{4ip}-1)(1+fg)$ to decapsulate 6-to-6 over 4 tunnels. Likewise, since the AR_h is a router, fragmentation constant $(1+fg)$ is considered. Each hop incurs process costs as much as $v_{6sec}v_{6ip}t_1$ during transmission to HA.

In Fig. 3, there are 4 tunnel processing places on the path between MR and HA. Its implication is that all the end points of each tunnel (i.e. MR, AR, f, AR_f, Ar-h, HA) should reflect their fragmentation constant.

$$V_{m2} = (2P_{rate}(C_{sh}(v_{6sec}-1)+C_{ip}v_{6sec}(v_{6ip}-1)(1+fg)+C_{ip}v_{6sec}v_{6ip}(v_{4ip}-1)(1+fg)) + v_{6sec}v_{6ip}v_{4ip}t_i + 2v_{6sec}v_{6ip}t_1)N_n$$

Fig. 4 shows that the Internet is composed of two carriers with two different versions. Beyond the boundary, IPv6 connectivity can be maintained. However, reverse tunneling including end-to-end VPN, is established on the MR to HA section.

$$V_{m3} = (P_{rate}C_{sh}(v_{6sec}-1)(2+fg)+2P_{rate}C_{ip}v_{6sec}(v_{6ip}-1)+2P_{rate}C_{ip}v_{6sec}v_{6ip}(v_{4ip}-1)(1+fg) + v_{6sec}v_{6ip}v_{4ip}(t_1+(t_i/2))+v_{6sec}v_{6ip}((t_i/2)+t_1))N_n$$

Fig. 5 refers to a model that can be established in the prior stages of all IPv6 transition phase environments, in which case MR can form connectivity through 6to6over4 tunnels up to the boundary facing the Internet.

$$V_{m4} = (P_{rate}(C_{sh}+C_{ip})(v_{6ip}v_{6sec}-1)(2+fg)+2P_{rate}C_{ip}v_{6sec}v_{6ip}(v_{4ip}-1)(1+fg) + v_{6sec}v_{6ip}v_{4ip}t_1 + v_{6sec}v_{6ip}t_i + v_{6sec}v_{6ip}t_1)N_n$$

Such models are considered as V_{6gtog} and V_{4gtog} as a reference point to compare influences of processing performance with those of NEMO VPN models. In addition, the costs incurred in gateway-to-gateway VPNs created between MR and HA were taken as a reference point. In the case of V_{4gtog} , even if it is not defined in the NEMO standard, the IPv4 VPN is compared in a cost analysis of the IPv6-based NEMO VPN in a transition environment.

4 Numerical Results

In the previous section, we have evaluated performance costs to set up connectivity in conjunction with VPN models according to each transition phase. In this section, factors affecting performance through numeric analysis for the NEMO VPN model under an IPv6 transition environment, are analyzed. As presented in Table 1, described values are applied to parameters that given are obtained from previous work.

Although it may raise the problematic from the standpoint of fairness because those parameters are not evaluated on the same environment, we use the parameter values in order to analyze relativity performance overhead for evaluation results. That is, we analyze relativity overhead for the processing performance of each NEMO VPN under a transition environment in contrast with V_{6gtog} and V_{4gtog} as a base line. In addition, it is assumed that a router or host has the same capabilities and resources for processing transmission packets.

The performances in VPN connectivity are analyzed according to increasing the amount of transmission packets from 64kbytes to several megabytes. Fig. 6 presents the differences in processing costs of VPN models with an increase in transmission packets from a single host of each model. In the case of V_{m2} , it can be seen that processing costs are highest, as are compared with that of V_{6basic} , V_{m1} , V_{m3} and V_{m4} presents differences not only in costs of reverse tunnels, including IPSec tunnels composed between MR and HA, bus also costs according to the length of such tunneling sections. The hierarchical tunnel means a tunnel within a tunnel, while the sequential tunnel means concatenated tunnels. Compared with the case of V_{4gtog} , V_{6gtog} , this suggests that NEMO VPN under transition creates 3 levels of hierarchical tunnels, and represents the reason costs have increased with an increase in transmission packets and hosts joining VPN, as shown in Fig. 7.

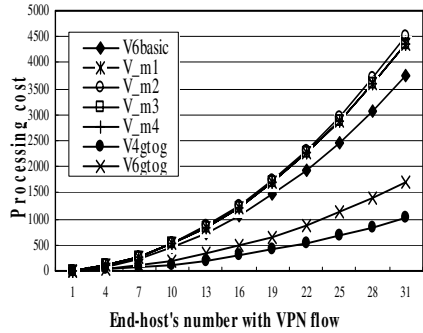
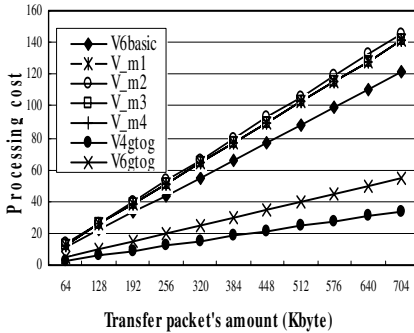


Fig. 6. VPN cost with an increase in packets Fig. 7. VPN cost with an increase in end-hosts

In NEMO VPN models, there is no difference in processing costs for an increase in packets due to the length of tunnel transmission and additional packet frequency. This implies that hierarchical tunnel processing causes a delay in transmission with heavy overhead. Fig. 8, demonstrates processing costs incurred under each model are influenced as the fragmentation constant increases. As mentioned in Section 4.2, it is assumed that these fragmentation constants represent the processing capability of a network device processing their packets. Therefore, when the results as shown in Fig. 6 and 7 are expected, the influences due to such processing capability of the network device can be analyzed.

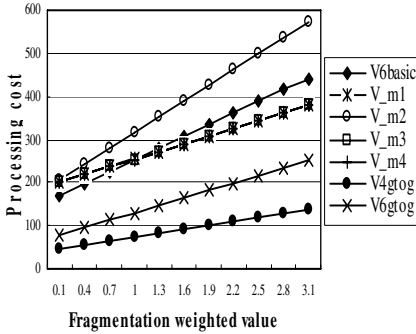


Fig. 8. VPN cost with an increase in fragmentation constant

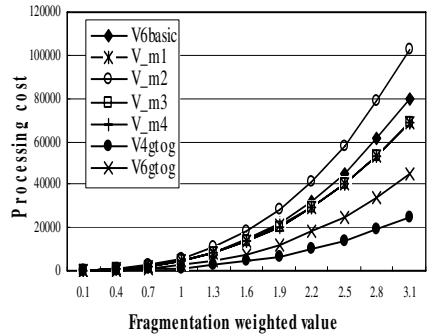


Fig. 9. VPN cost with an increase in fragmentation constant and and hosts

Fig. 8 explains that, as fragmentation constants increase, models such as V_{4gtog} , V_{6gtog} incur a lower increase in processing costs than NEMO VPNs created in 3 levels of hierarchical tunnels. That is, NEMO VPN models have presented a greater increase due to such hierarchical tunnels. Especially, with regard to the location of tunnel processing, V_{6basic} must cover 2 tunnel end points; V_{m1} V_{m3} V_{m4} 3 tunnel end points; V_{m2} 4 tunnel end points. In this case, it can be seen that influences occurring upon processing costs according to the processing capabilities of network devices. Like Fig. 7, Fig. 9 demonstrates how processing costs are influenced when the frequency of processing packets increases with regard to the number of hosts joining the VPN.

In V_{m2} with the largest number of processing locations, processing costs increase relative to existing levels. In Fig 8, V_{6basic} incurs more costs over another NEMO VPN model, when fg increases. These results demonstrate that, because the rate in which v_{6ip} increases is higher than that of v_{4ip} , it has more influence upon the costs of fragmentation according to the processing capability of the network device.

5 Conclusion

The current network mobility base specification requires that all signaling messages between the MR and the HA must be authenticated by IPsec. The use of IPsec to protect Mobile IPv6 signaling messages is described in detail in the HA-MN IPsec specification by the IETF. In order to protect against a wide range of attacks, using AH and/or ESP between MR and HA is of paramount importance. The purpose of this study is to analyze relative influences of VPN tunnel and IP tunnel upon processing capabilities, based on cryptographic processes in NEMO topology. In particularly, we analyzed VPN scenarios applicable to mobile network environments during IPv6 transition phases. According to the network environment, various NGtrans mechanisms are used for communication with the IPv4 node.

The correlation between transition mechanisms used in the existing network and transition mechanisms supported by a mobile network can become an important decision factor in establishing secure connectivity with the home network. As discussed previously, there are restrictions in applying VPN in conjunction with transition mechanisms, because there are fundamentally conflicting transition mechanisms and IPsec VPN; the former supports connectivity between different IP versions by modifying the address and format of IP packets, the latter supports the security service by prohibiting the modification of transit packets. Alternatively, hierarchical tunnel-based transitioning solutions may be used when security is a requisite. However, encapsulating IPv6 packets within IPsec payloads, and then within IPv4 packets, results in significant overhead, therefore resulting in degraded performance.

In conclusion, as presented in the results, the NEMO VPN composing of hierarchical or sequential tunnels should be proposed only after careful evaluation of security vulnerabilities and performance requirements. Most vendors have been developing IPv6 in the same aspect with our study plan in IPv6-based packet forwarding technologies based on hardware. Likewise, when it comes to the NEMO community, if the configuration of both IP tunnel and VPN tunnel on mobile routers as well as hosts is required under an IPv6 transition environment, the relative influences upon processing capabilities should be considered sufficient. Incidentally, it is recommendable that the future NEMO specification fully reflects how it can be formed from a VPN performance perspective. Also, in relation to multiple tunnels, SA establishment between tunnel end points, reliable tunnel configuration, and IPsec processing burden represents an important problem to overcome for successful wide deployment of mobile networks.

References

1. Gilligan, R. and E. Nordmark, "Transition Mechanisms for IPv6 Hosts and Routers", RFC2893, August 2000.
2. Thierry Ernst, "Network Mobility Support Requirement," work in progress, available at draft-ietf-nemo-requirements.txt, February 2003.
3. Thierry Ernst, "Network mobility in IPv6", PhD Thesis, University Joseph Fourier Grenoble, France, available at <http://www.inria.fr/rrrt/tu-0714.html>, October 2001.
4. Thierry Ernst, "Network Mobility Support Terminology", work in progress, available at draft-ernst-nemo-terminology.txt, November 2002.
5. P. Thubert, M. Molteni, P. Wetterwald, "IPv4 traversal for MIPv6 based Mobile Routers", work in progress, available at draft-thubert-nemo-ipv4-traversal-01.txt, May 2003.
6. S. Zeadally, R. Wasseem, I. Raicu, "Comparison of End-System IPv6 Protocol Stacks", IEE Proceedings Communications, Special issue on Internet Protocols, Technology and Applications (VoIP), Vol. 151, No. 3, June 2004.
7. Raicu and S. Zeadally, "Impact of IPV6 on End-user Applications", IEE/IEEE International Conference on Telecommunications (ICT 2003), Tahiti, Papeete, French Polynesia, February 2003.
8. E. Kandasamy, G. Kurup, T. Yamazaki, "Application Performance Analysis in Transition Mechanism from IPv4 to IPv6", APAN Conference, Tsukuba, February 2000.

9. I. Raicu and S. Zeadally, "Evaluating IPv4 to IPv6 Transition Mechanisms", IEE/IEEE International Conference on Telecommunications (ICT 2003), Tahiti, Papeete, French Polynesia, February 2003.
10. Seiji Ariga, Kengo Nagahashi, Masaki Minami, Hiroshi Esaki, Jun Murai, "Performance Evaluation of Data Transmission using IPSec over IPv6 Networks", INET2000, Yokohama, July 2000.
11. Hyung-Jin. Lim, Yun-Ju Kwon, Tai-Myoung. Chung, "Secure VPN Performance in IP Layers", The Journal of The Korean Institute of Communication Sciences Vol. 26 No. 11, 2001.
12. Mogul, Jeffery, and Stephen Deering, "Path MTU Discovery". RFC 1191, November 1990.

A Short-Lived Key Selection Approach to Authenticate Data Origin of Multimedia Stream^{*}

Namhi Kang¹ and Younghan Kim²

¹ Ubiquitous Network Research Center in Dasan Networks Inc.,
Yatap-Dong, Bundang-Ku, 463-070 Seongnam, South Korea
nalnal@dcn.ssu.ac.kr

² Soongsil University, School of Electronic Engineering,
Sangdo 5-Dong 1-1, Dongjak-Ku, 156-743 Seoul, South Korea
ykim@dcn.ssu.ac.kr

Abstract. This paper presents an efficient approach to provide data origin authentication service with a multimedia stream application. The proposed approach is intended to achieve not only fast signing/verifying transaction but also low transmission overhead. In particular, the consideration on the performance is one of key issues for increasingly widespread using of wireless communication since there are lots of limitations including scarce network resources, low computing power and limited energy of nodes. To meet such requirements, we take advantage of using a short-lived key(s) that allow an authentication system to overcome the performance degradation caused by applying highly expensive cryptographic primitives such as digital signature. The major concern of this paper, therefore, is to derive an appropriate length of a key and hash value without compromising the security of an authentication system.

1 Introduction

Multimedia streaming technology is already prevailing throughout the Internet. Multicast is regarded as a cost efficient delivery mechanism for multimedia streaming from a source to a group of receivers. However, in contrast to regular unicast communications, multicast is more difficult and has several inherent obstacles that still remain to be solved [1]. This paper is intended to find a way to support data origin authentication for multimedia multicast applications. We focus on the performance consideration that is necessary for both wireline and wireless communication. Especially, computational and transmission overheads are much concern in the lossy and unreliable wireless communication.

Data origin authentication (DOA) is commonly achieved by MAC (Message Authentication Code) in unicast [2]. However, it is difficult to apply MAC directly into multimedia streaming over multicast channel owing to a couple of challenges, such as the difficulty of key sharing and the possibility of insider

^{*} This work was supported by the Korea Research Foundation Grant. (KRF-2004-005-D00147).

coalition to cheat other members. Applying digital signatures instead of MAC can solve these shortcomings since only the source is able to bind its identity to the signature. The trade-off is nevertheless that there exist critical performance problems when an asymmetric cryptography primitive is employed to a multimedia stream.

Such problems of most regular digital signature schemes such as RSA and DSA are caused by using a large key size for protecting data in the context of long-term security. It is too ambiguous however to specify how much time is enough for the long-term security. The validity term of security obviously differs depending on the application, the policy or the underlying environment. Multimedia streaming applications have different requirements and properties from the viewpoint of security service. If an application is required to support authentication in order to resist against attacks only for a short period such as a week or a day, then using a key which is long enough for one or two decades is a waste of resources. The best exemplified application is the entertainment contents streaming service (e.g. movies or music), where the recipient discards a received packet as soon as it is consumed. This is widely used scenario in multimedia streaming applications.

On the basis of this notion, we propose an efficient approach to support DOA for multimedia streaming using a short-lived but secure key. The proposed approach is not a new authentication scheme in itself, but a framework to employ existing schemes based on digital signature (see section 2). The proposed approach is able to offer the equivalent security with smaller key sizes.

This paper is organized as follows. We discuss stream authentication solutions proposed so far in the literature in section 2. In section 3, the lower bound of the Lenstra and Verheul's suggestion [3] that is the basic theory of this paper is shortly described. In section 4, we propose the way to select a short-lived key in a secure fashion. In section 5, we report some comparison results, and then we conclude this paper in section 6.

2 Multicast Stream Authentication Solutions

Existing streaming authentication solutions that we consider in this paper are divided into two categories: MAC based approaches [4, 5] and digital signature based approaches [4, 6, 7, 8, 9, 10]. In this paper, TESLA [4], WL' Tree and Star scheme [7], EMSS [4], and SAIDA [9] are discussed and compared in detail.

Multicast data origin authentication is intended for all receivers to ensure that the received data is coming from the exact source rather than from any member of the group. The latter is called group authentication in multicast [5]. Therefore, the guarantee of *asymmetric property*, which means only the signer (the source) can generate authentication information while others (receivers or verifiers) is only able to verify such information, is required. To achieve this asymmetric property, TESLA is very efficient since it requires only three MCA computations per packet resulting also in a low transmission overhead. However, TESLA requires the time synchronization between the source and the receiver at

the initial bootstrapping time. On the other hand, the limitation of the Multiple-MAC scheme is that there still exist attacks caused by an insider coalition.

In digital signature based schemes, asymmetry is a natural property since it uses an asymmetric key pair: a private key for the signer and a public key for the verifier. Most solutions based on a digital signature employ an amortizing signature over several packets [4, 7, 8, 9, 10]. In these schemes, time expensive signing and verifying operations are performed once a block which includes a set of packets. There are two conspicuous differences between them: one is the method of setting a group of packets for amortizing, and the other is the way to achieve the loss tolerant property.

With a system applying an amortizing signature, there is a trade-off between the computational cost and the verification delay. As the block size increases, where the block size is the number of packets gathered to set a block for amortizing, the computational time is decreased, whereas the verification delay is increased owing to the waiting time including both for generating a block at the sender and for the arrival of a signature at the receiver side. In addition, packet loss is a critical issue when an authentication system applies an amortizing signature over stream.

Wong and Lam proposed star and tree technique (referred to as the WL scheme in this paper) [7] based on Merkle's hash tree technique [11]. The basic idea of the WL scheme is to sign a block hash which is the top level hash value of the tree representing all packet's hash in a block so that a signature (called block signature) amortizes all packets in a block. Hash chaining was introduced in [6], where only the first packet including hash of the next packet is digitally signed. Thereafter, the source continuously sends each packet which contains the hash value of its next packet. This scheme is efficient but it is not loss tolerant and the source has to know the entire stream in advance. In order to overcome these shortcomings, several solutions (e.g. [4] and [8]) have been proposed recently. An erasure coding function can be used to achieve space efficiency in streaming authentication under the condition that the predefined number of packets must be arrived at the receiver [9, 10]. The drawback of those schemes is more expensive computational cost caused by applying erasure code than other schemes.

3 Overview of Lenstra and Verheul's Suggestion

In the proposed approach, the source is intended to support DOA without high performance degradation of a multimedia application. To do so, the source generates a signature using a short key of length which is available only for a predefined short duration. The challenge is to derive an appropriate length of key without compromising the security of an authentication system. Our approach is based on the lower bound of the Lenstra and Verheul's suggestion [3].

In [3], Lenstra and Verheul recommended appropriate key sizes for both symmetric and asymmetric cryptosystems based on a set of explicitly formulated parameter settings combined with historic data points of attacking on the

cryptosystems. In particular, the computational efforts involved in existing successful attacks, namely a number of MIPS (Million Instructions Per Second) years required for a successful attack, are the major concern.

They defined the $IMY(y)$ as a MIPS-Years considered to be infeasible until year y to derive key size necessary to support computationally equivalent to the same strength offered by the DES in 1982. The computational efforts of 5×10^5 Mips-Year is considered as secure cost to resist software attack on commercial DES in 1982. $IMY(y)$ is formulated by

$$IMY(y) = 5 * 10^5 * 2^{12(y-s)/m} * 2^{(y-s)/b} \text{ MIPS-Years}, \quad (1)$$

where s , m , and b are variables to consider environmental factors affecting the security in selection of key length (see [3] in detail).

On the basis of equation (1), the lower bound of key size for a symmetric key system (also for a hash function) is derived as the smallest integer that is at least

$$56 + \log_2(IMY(y))/5 * 10^5. \quad (2)$$

On the other hand, to select the length of the conventional asymmetric key system, they applied the asymptotic run time $L[n]$ of a NFS (Number Field Sieve) combined with a historical fact that a 512 bits modulus was broken in 1999 at the cost of around 10^4 MIPS year, such that

$$\frac{L[2^k]}{IMY(y) * 2^{12(y-1999)/18}} \geq \frac{L[2^{512}]}{10^4}, \quad (3)$$

where a factor $2^{12(y-1999)/18}$ indicates that the cryptanalytic is expected to become twice as effective every 18 months during year 1999 to year y .

4 Our Approach

In [12], we proposed an approach to support secure routing methodology for mobile nodes, where a secure length of key pairs for a digital signature was derived. Yet no consideration on multimedia streaming authentication and cryptographic hash function has previously appeared. In this section, we briefly review the way presented in [12] and extend it for finding a secure length of hash function.

4.1 Secure Key Length Selection for Digital Signature

The $IAO(x, y)$ was defined in [12] as an ‘Infeasible Amount of Operations’ within a certain period x of the year y . $IAO(x, y)$ is derived from the $IMY(y)$ formulated in equation (1). $IAO(1 \text{ year}, \text{ this year})$ is therefore derived from $IMY(\text{this year})$. Here, it is emphasized that the term ‘MIPS’ is used as a measuring tool for the relative comparison of computational workload. Like [3], it is also supposed that a single operation requires a single clock cycle. That is, a 2.4 GHz Pentium 4 computer (which is used for our experiments) is regarded as a 2400 MIPS machine. A short but secure length of key was calculated based on the following intuitive corollary.

Corollary 1. Let $S = \{x_1, \dots, x_j : x_1 + \dots + x_j = X\}$ be a set of sub periods of X , and k be the length of a short-lived key. If k meets equation (4), then the length of k is secure enough within x_i of the year y , for all $x_i \in S$.

$$L[2^k] \geq 55.6 * 2^{12(y-1999)/18} * IAO(x_i, y). \tag{4}$$

Proof. We derived equation (4) by applying $IAO(X, y)$ to equation (3), where we approximated that $\frac{L[2^{512}]}{10^4}$ is equivalent to $1.75 * 10^{15}$. The proof was appeared in our previous work (see [12] in detail). \square

Example: Consider 2004 as the target year (y of $IMY(y)$) to protect a system, for example, then it is expected that $IMY(2004)(= 5.98 * 10^{10}$ MIPS) is secure until the end of year 2004 according to equation (1). This is equivalent to $1.88 * 10^{24}$ operations thus such a number of operations are infeasible until the end of 2004, namely $IAO(1 \text{ year}, 2004)$ is $1.88 * 10^{24}$. Thereby we can calculate a proper key length for any sub period as described in Table 1. If a day is determined as a short period, the source can use an 880-bit key to sign. This is because $IAO(1 \text{ day}, 2004)$ is $5.16 * 10^{21}$ which is equivalent to $1.64 * 10^8$ MIPS year. 880-bit can be derived by using equation (4). The signature computed with an 880-bit key offers the same security level for a day as does a 1108-bit key for this year and a 2054-bit key for the next two decades.

Fig. 1 shows that the cost of a general digital signature scheme depends on the key length, whereas our approach depends on the lifetime of the deriving short-lived key.

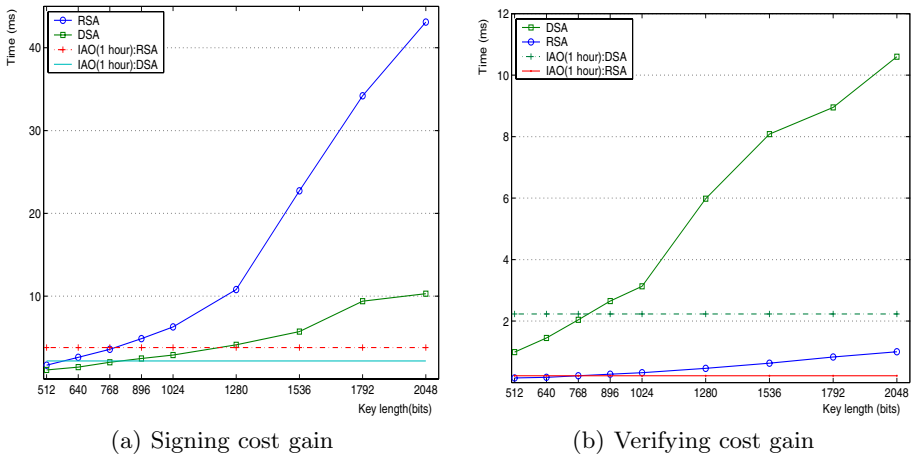


Fig. 1. Computational cost gain (experiment conditions are same with table 1)

In DSA, it is known that the time it takes to generate a signature is lower than that of signature verification, but the signing cost is similar to verifying cost in our experiment (see Table 1.). Unlike DSA, signature verification is much

faster than signature generation in a RSA setting. It might be claimed that it is advantageous for signing to be the faster operation. However, it may well be more advantageous to have faster verification in multicast, especially one-to-many applications. This is because, in the usual scenario, a content distributor has much higher capabilities, such as fast CPU and large RAM, than those of the receivers. For this reason, we consider RSA as an underlying signature algorithm in this paper.

4.2 Hash Length Selection

Like [10], we also apply a TCR (Target Collision Resistance) hash function into an authentication scheme instead of ACR (Any Collision Resistance) under the assumption that there exist UOWHFs (Universal One-Way Hash Functions). The TCR hash function allows the authentication scheme using amortizing signature to gain much more space efficiency. The major advantage of TCR compared with ACR is that the birthday attack, which is the best attack on ACR hash function families, does not directly apply to TCR hash function since message is specified before the hash function is given an attacker. Unlike TCR, an attacker can select two different messages, M and M' , that map to the same hash, $h(M) = h(M')$, in advance with a setting of ACR hash function. As a result, TCR hash function of size $L/2$ can satisfy with the same secure level to ACR hash function of size L .

For example, if an authentication system has a target period of less than 18.75 hours as the x_i of $IAO(x_i, y)$, then only an 8 byte hash value is sufficient to hold TCR. In particular, the transmission overhead shortcoming of WL's Tree scheme, where $\log_2 BS$ hashes per packet are required for authentication, can be reduced by a factor of more than 2. Furthermore, from the viewpoint of computational cost, the source can use a 128-bit hash function (e.g. MD5) for all periods of $IAO(x_i, y)$ in the case of applying UOWHF, while the source must use a 160-bit or higher-bit hash function (e.g. SHA1, RIPEMD160, or SHA256) for a period of more than a day in order to protect against birthday attack in an ACR hash function setting.

A secure hash length for the predefined lifetime is calculated based on the following theorem 2. If an authentication scheme does not hold TCR property, an authentication system must replace TCR with ACR hash function [10]. With a system applying TCR hash function, for example, there is a secure hole that the source signs on a malicious packet which is a counter part of a collision pair. It is possible because the source knows the TCR hash function to be used in advance. However, this is not the case of DOA but the case of relating to NRO. We recall again that NRO is not addressed in our scenarios, as discussed above.

Corollary 2. *Let $S = \{x_1, \dots, x_j : x_1 + \dots + x_j = X\}$ be a set of sub periods of X , and l_{TCR} be the hash length of TCR hash function (generally $l_{TCR} = l_{ACR} * 1/2$). If l_{TCR} meets equation (5), then the length of l_{TCR} is secure enough within x_i of the year y , for all $x_i \in S$.*

$$l_{TCR} \geq 56 + \log_2\left(\frac{IAO(x_i, y)}{15.75 * 10^{18}}\right). \quad (5)$$

Table 1. Short-lived key and hash length selection including computational cost

Period x_i	$IAO(x_i, y)$ ($y = 2004$)	Key length (bits)	RSA (msec) sign/verify	DSA (msec) sign/verify	Hash(bits) ACR/TCR
2 decades	$7.78 * 10^{28}$	2113	51.55 / 1.13	12.96 / 13.55	178 / 89
1 decades	$3.81 * 10^{26}$	1562	29.31 / 0.65	6.13 / 9.31	162 / 81
1 year	$1.88 * 10^{24}$	1108	7.92 / 0.31	3.64 / 4.15	146 / 73
1 Month	$1.57 * 10^{23}$	1008	5.92 / 0.29	2.87 / 3.14	140 / 70
1 Week	$3.67 * 10^{22}$	952	5.37 / 0.27	2.71 / 2.82	136 / 68
1 Day	$5.15 * 10^{21}$	880	4.55 / 0.23	2.34 / 2.71	130 / 65
1 Hour	$2.15 * 10^{20}$	772	3.79 / 0.22	2.18 / 2.23	122 / 61
30 minutes	$1.07 * 10^{20}$	748	3.27 / 0.20	1.93 / 2.0	116 / 58

Proof. Corollary 2 follows the proof of corollary 1. Equation (5) is derived by applying $IAO(x_i, y)$ to equation (2) in a similar way to calculate equation (4) in corollary 1. \square

In Table 1, we tabulate the length of signature key and hash for each selected period (lifetime) as well as computational cost when the selected key is applied to an authentication solution. To experiment such computation time, we use Crypto++ library [14] on a Linux based 2.4GHz Pentium 4 computer, where the input size is 1024 bytes.

4.3 Remark on Short-Key Management

The source may have to generate several ephemeral key pairs during the session according to the validity term (say lifetime) of selected short key pairs. The lifetime available for secure use is proportional to the strength of the key being used such as the length of RSA modulus. In some scenarios, multiple key pairs for a single session may lead to public key distribution and certification issues since the source should re-key and re-certify the newly distributed key. In particular such difficulty is even more challenging in a large group. However, it is not the case of the proposed approach if the key selection is appropriate according to the property of an application. We note that the time required for most multimedia contents distribution is not long (e.g. 2 hours for a movie, a quarter or half a day for an on-line presentation). In such a scenario, a key pair for a day, namely $IAO(1 \text{ day}, y)$, can be a good selection without the overburden necessary to redistribute a new public key securely. In this paper, we do not address the issue of key management in detail. Instead we assume that every recipient has ascertained the correctness of a public key corresponding to a private key (secret key) used for packet signing operations in advance by use of key management protocol and public key certification mechanisms such as

PKI (Public Key Infrastructure) or GDOI (Group Domain Of Interpretation) which can be especially used when no PKI is available [13].

5 Performance Evaluation

We evaluate performance efficiencies to compare our approach with four other schemes (i.e. WL’s Tree, EMSS, SAIDA and TESLA that are briefly described in section 2) in terms of computational cost as well as the transmission overhead. We commonly used 1024-bytes as the length of input data to get the computational time of each primitive, and experiments were performed in the same conditions as those of Table 1, namely, 1024 byte input, Crypto++ library on a 2.4GHz machine. To do an evaluation, a decade and a day were set as the security level (say the lifetime of a key) for existing schemes and our approach, respectively. In the case of SAIDA, we assumed up to 50% losses per block. In addition, we set 6 edges for EMSS scheme (namely, 6 hashes per packet).

Fig. 2 shows that the proposed approach is very efficient with respect to computational cost at the source side (see (a) in Fig.2) and at the receiver side (see (b) in Fig.2) as well, resulting from the use of a short-lived key. From the viewpoint of computational cost, SAIDA is more expensive than others using an amortizing signature owing to the additional cost for applying an erasure coding function. EMSS and WL’s Tree require similar computational cost. In particular, (a) of Fig 3 shows that our approach applied to EMSS and WL’s Tree schemes require lower computational cost than even TESLA, which uses fast MAC as the underlying primitive, in the case where group members are widely dispersed throughout the Internet so that each packet experiences a different delay in

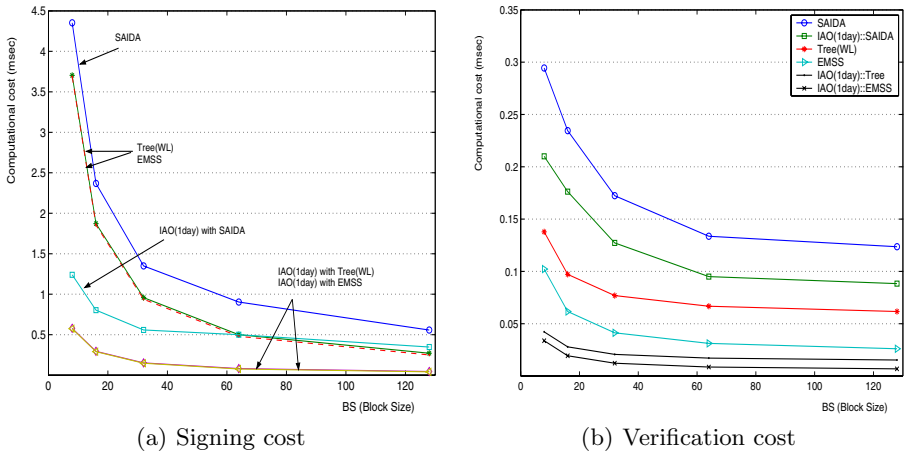


Fig. 2. Computational cost comparison

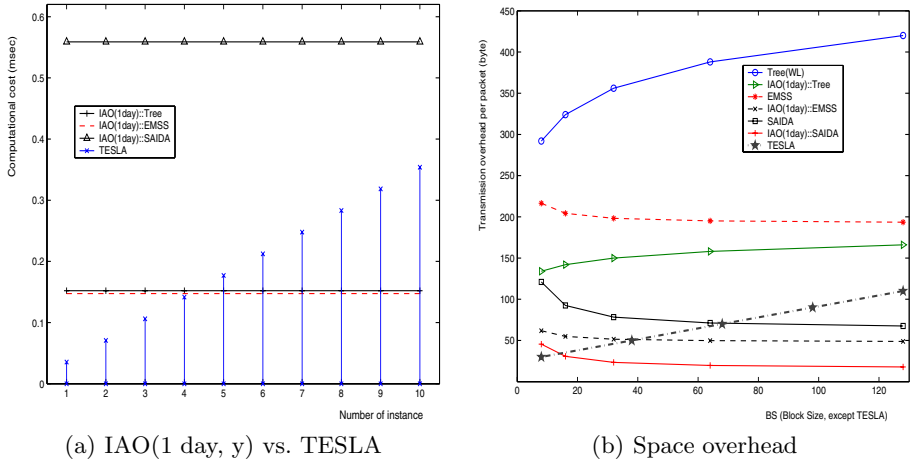


Fig. 3. Computing and space overhead comparison

transit. To compensate for the different arrival time in a heterogeneous multicast network, TESLA uses multiple key chains; the source calculates several MACs for a packet with different keys and each key will be sent with a different delaying lag. The number of instance (x -axis of (a) in Fig. 3) indicates the number of different delay groups.

In Fig. 3, (b) illustrates the space overhead of each solution versus the number of packets per block, where TESLA (indicated by a dotted line with star point) is exception. In the case of TESLA, each point on the line shows the effect of the number of instances on the space overhead (2,4,6,8, and 10 respectively). From the aspect of the transmission overhead, unlike the comparison of computational cost, SAIDA is the best. The Tree scheme is likely to be limited for use in practice due to a high overhead. However, the Tree scheme allows the receiver to verify a packet immediately and individually, regardless of the loss pattern. In the case of applying IAO to Tree, the space overhead is reduced by a factor of 2 at a minimum. As shown in the figure, EMSS and SAIDA with $IAO(1\ day, y)$ become more space efficient than TESLA in the scenario where there are more than four delay groups in multicast and the source amortizes a signature over more than 40 packets.

6 Conclusion

In this paper, we first discussed a couple of existing solutions to support data origin authentication for multicast stream. Then we proposed an efficient approach using a short-lived key. To do so, we also addressed how to calculate an appropriate length of key and hash value for a predefined short lifetime. The results show that our approach is a very efficient way to enhance the performance of stream authentication solutions.

References

1. C. Diot, B. Levine, B. Lyles, H. Kassem and D. Balensiefen, "Deployment Issues for the IP Multicast Service and Architecture," IEEE Network magazine, January/February 2000.
2. S. Kent and R. Atkinson. Security Architecture for the Internet Protocol. IETF RFC2401, Nov. 1998.
3. Arjen K. Lenstra and Eric R. Verheul. Selecting cryptographic key sizes. Journal of Cryptology 14(4):255-293, 2001.
4. A. Perrig, R. Canetti, J. D. Tygar and D. Song. Efficient Authentication and Signing of Multicast Streams over Lossy Channels. IEEE Security and Privacy Symposium, 2000.
5. R. Canetti, J. Garay, G. Itkis, D. Micciancio, M. Naor and B. Pinkas. Multicast Security: A Taxonomy and Some Efficient Constructions. Infocom'99, 1999.
6. R. Gennaro and P. Rohatgi. How to Sign Digital Streams. Lecture Notes in Computer Science, vol. 1294, pages 180-197, 1997.
7. C. K. Wong and S. S. Lam. Digital Signatures for Flows and Multicasts. In Proc. IEEE ICNP '98, 1998.
8. P. Golle and N. Modadugu. Authenticating streamed data in the presence of random packet loss. NDSS'01, pages 13-22, Feb. 2001.
9. Jung Min Park and Edwin K. P. Chong. Efficient multicast stream authentication using erasure codes ACM Trans. Inf. Syst. Secur. 6(2):258-285, 2003.
10. C. Karlof, N. Sastry, Y. Li, A. Perrig, and J.D. Tygar. Distillation Codes and Applications to DoS Resistant Multicast Authentication. NDSS 2004.
11. R. Merkle. Protocols for public key cryptosystems. In Proceedings of the IEEE Symposium on Research in Security and Privacy, pages 122-134, Apr. 1980.
12. N. Kang, I. Park, and Y. Kim. Secure and Scalable Routing Protocol for Mobile Ad-hoc Networks, Lecture Notes in Computer Science, vol 3744, Oct. 2005.
13. Baugher, M., Weis, B., Hardjono, T. and H. Harney, The Group Domain of Interpretation, RFC 3547, July 2003.
14. Crypto++ class library, <http://www.eskimo.com/~weidai/cryptlib.html>

Weakest Link Attack on Single Sign-On and Its Case in SAML V2.0 Web SSO

Yuen-Yan Chan

Department of Information Engineering,
The Chinese University of Hong Kong,
Shatin, N.T., Hong Kong
yychan@ie.cuhk.edu.hk

Abstract. In many of the single sign-on (SSO) specifications that support multitiered authentication, it is not mandatory to include the authentication context in a signed response. This can be exploited by the adversaries to launch a new kind of attack specific to SSO systems. In this paper, we propose the *Weakest Link Attack*, which is a kind of parallel session attack feasible in the above settings. Our attack enables adversaries to succeed at all levels of authentication associate to the victim user by breaking only at the weakest one. We present a detailed case study of our attack on web SSO as specified in Security Assertions Markup Language (SAML) V2.0, an OASIS standard released in March, 2005. We also suggest the corresponding repair at the end of the paper.¹

1 Introduction

Authentication is almost unavoidable in performing business processes over the Internet. Very often, users need to sign-on to multiple systems within one transaction. What makes it worse is that different credentials and authentication methods are often involved. System administrators also face the challenge of managing the increasing number of user accounts and passwords. One solution for this issue is *Single Sign-On (SSO)*, a technique that enables a user to authenticate once and gain access to the resources of multiple systems. In order to facilitate SSO, enterprises unite to form circles of trust, or *federations*. They do so by establishing business agreements, cryptographic trust, and user identities across security and policy domains. Nowadays, federation and SSO become the dominant movement in identity management in electronic business.

To this end, standards and specifications are published by renown organizations to provide standardized mechanisms and formats for the communication of identity information within and across federations. The most influential one is the Security Assertion Markup Language (SAML) established by the Organization for Advancement of Structured Information Standards (OASIS), which is an XML-based framework for communicating user authentication and attribute

¹ This piece of work is financially supported by Hong Kong Earmarked Grants 4232-03E and 4328-02E of Hong Kong Research Grant Council.

information. Its first version, SAML V1.0 [2], was published in November, 2002. Since then, OASIS has released subsequent versions of SAML and the latest one is SAML V2.0 [3] that released in March, 2005. Today, SAML has been broadly implemented in major enterprise Web server and application server products. Examples include BEA WebLogic ServerTM9.0, Sun Java System Application Server PlatformTMEdition 8.1, IBM WebSphere Application ServerTM6.0, as well as commercial products for SSO and identity federation solutions such as Computer Associates eTrust Single Sign-OnTM and IBM Tivoli Federated Identity ManagerTM.

1.1 Related Work

Related works on security analysis of SSO include the following. In the OASIS Standard [6], a number of threats to the SAML V2.0 web SSO specification are documented, in which countermeasures that heavily depend on SSL/TLS as well as system level mechanisms are highlighted. These threats include message eavesdropping, replay and modification, man-in-the-middle attacks, and denial of service attacks. As of the completion of this paper, there are no other attacks reported on this latest SAML version yet. Groß analyzed the security of SAML V1.1 SSO Browser/Artifact Profile [1] and revealed three attacks, namely the connection hijacking attack, man-in-the-middle attack, and the HTTP referrer attack. This caused OASIS SSTC² to respond with a document recently [7] and restate the security scope of SAML. A few work with a different approach, such as Hansen *et. al.* who validated SAML V1.1 SSO security using static analysis and reported several flaws in some SSO instantiations [9]. Some work has also been done on security analysis of Liberty SSO. For example, Pfitzmann *et. al.* presented a man-in-the-middle attack of Liberty ID-FF V1.1 SSO with enabled client [8]. Liberty corrected the flaw in ID-FF V1.1 with the countermeasure proposed by the authors.

1.2 Our Contribution

In this paper, we present the *weakest link attack*, which is a parallel session attack specific to SSO systems. We start by an attack hypothesized in a generic SSO scenario where multitiered authentication is in place. Then we examine the attack hypothesis by exploiting it on SAML V2.0 web SSO and we succeed. We further generalize our attack and propose a repair for its elimination.

Since OASIS SSTC position SAML as a common XML framework, and emphasize that SAML protocols and implementations are designed to operate in a broader context in conjunction with other protocols and mechanisms [6], we do not regard we have broken SSO in SAML and its related specifications. Nevertheless our paper achieves significant attack results and contributes valuable information and considerations to future implementation of SSO in general.

² Security Services Technical Committee.

1.3 Paper Organization

After an overview of SSO and SAML in Section 2, the rest of the paper is arranged in a four-step style:

1. *Attack Hypothesis*. We will provide the definition and the hypothesis of the attack in Section 3.
2. *Attack Exploitation*. We will exploit the hypothetical attack on a default instantiation of SAML V2.0 web SSO in Section 4.
3. *Attack Generalization*. Based on the results obtained, we will generalize the attack over generic SSO scenarios in Section 5.
4. *Attack Elimination*. Lastly, repair to the flaw in the SSO specification will be provided in Section 6.

and the paper will be concluded in Section 7.

2 An Overview on Single Sign-On and SAML

2.1 Single Sign-On

Single Sign-On is mechanism which enables a user to authenticate virtually once and gain access to the restricted resources of multiple systems. Typically SSO involves three principals:

- *User U*. Who accesses the network and makes use of the system resources for any purposes. In practice, the user is often represented by a *user agent*.
- *Service provider SP*. Who offers restricted services which are only available to authenticated principals.
- *Identity provider IdP*. Who creates, maintains, and manages identity information for principals and provides principal authentication to other service providers.

We depict a generic SSO scenario in Fig. 1.

2.2 Security Assertions Markup Language (SAML)

The SAML Standard consists of a set of XML schemas and specifications, which together define how to construct, interchange, interpret, and extend security assertions for a variety of purposes. The major ones include web Single Sign-On (web SSO), identity federation, and attribute-based authorization. The Standard defines a set of assertions, protocols, bindings, and profiles [3]. An assertion is a piece of information that provide one or more statements made by an SAML authority, and SAML protocols specify how to generate messages and exchange them between requesters and responders. SAML protocol bindings [4] are mappings from SAML message exchanges into standard communication protocols such as SOAP and HTTP. The SAML profiles [5] define a set of rules to use SAML assertions and protocols to accomplish specific purposes, in particular, the Web Browser SSO Profile defines how web SSO is supported.

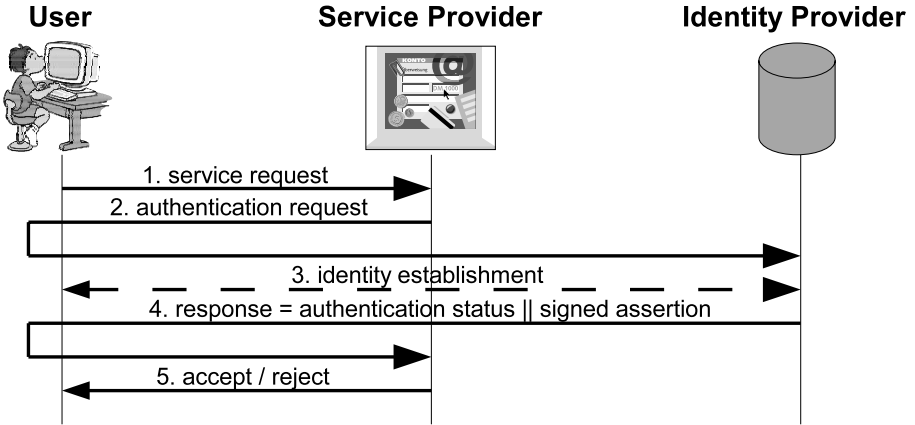


Fig. 1. Single Sign-On

2.3 SAML V2.0 Web SSO

A default instantiation³ of web SSO according to the specification in SAML V2.0 Web Browser SSO Profile [5] is illustrated in Fig. 2.

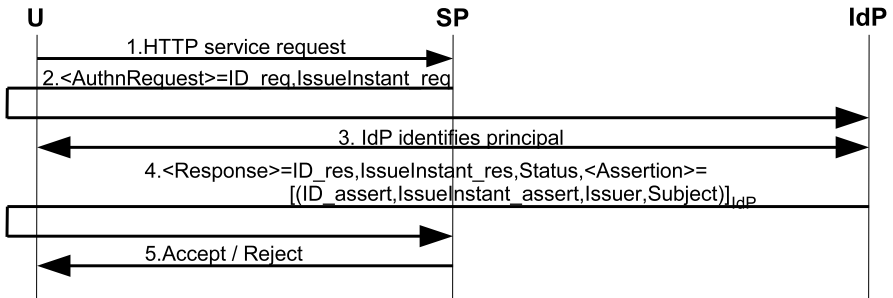


Fig. 2. SAML web SSO

2.4 Basic Security Measures of SAML V2.0 web SSO

Security measures in SAML Web Browser SSO Profile include [6] the use of timestamps (*IssueInstant*) to indicate message freshness, the requirement of IdP’s digital signature on *<Assertion>* to prevent forged assertions, the use of unique message identifiers (*ID*) to ensure one-use property of messages and assertions, and the requirement of digital signature on assertions by identity provider to ensure integrity.

³ This refers to an instantiation enforcing mandatory (i.e. [Required]) elements and attributes only.

3 Attack Hypothesis

Definition 1 (Weakest Link Attack). *A weakest link attack is a parallel session attack in single sign-on settings where an adversary launches concurrent service requests at two service providers that require different authentication levels, only at one of which the adversary can pass. The adversary makes use of the response corresponds to the successful authentication issued by the identity provider to replace that of the failing one, so that both requests are accepted.*

Consider an adversary Adv who has broken the level- \mathcal{L} authentication context of a user U. Under SSL/TLS, the power of Adv is confined to:

1. Authenticate as U at level \mathcal{L} .
2. Initiate arbitrary number of concurrent conversations with any other legitimate principals.
3. Obtain messages which U is the legitimate receiver.
4. Send messages which U is the legitimate sender.
5. Obtain and redirect messages which U is the legitimate redirector.
6. Alter plaintext messages which U is the legitimate receiver or redirector.

Hypothesis 1 (Weakest Link Attack). *Adversary Adv can succeed in the weakest link attack provided all of the following conditions are satisfied:*

- C1. There exists two or more service providers relying on a single identity provider for user authentication.*
- C2. Multitiered authentication is in place.*
- C3. Either one or both of the following conditions are true:*
 - a. The level of authentication or/and the designated service provider is/are not indicated in the response.*
 - b. There is no integrity for the response message.*
- C4. User redirection for request and response messages is required.*

Remark 1. The first condition is common in SSO practices, especially when the service providers are providing complementary services. E.g. an airline company and a car rental service company sharing a common travelers' club as their identity provider. The second condition is reasonable as requirements of authentication contexts diverge among service providers.

3.1 Weakest Link Attack in Hypothetic SSO

We describe how the hypothetic attack (Hypothesis 1) takes place using the SSO depicted in Fig. 1 as an hypothetic scenario. We assume the following settings: consider Adv to be an adversary described above. There are two service providers, SP1 and SP2, who both rely on the same identity provider IdP for authentication and assertion services. Suppose SP1 requires an authentication context at level- \mathcal{H} and SP2 requires that at level- \mathcal{L} where $\mathcal{H} > \mathcal{L}$ (for example, assume \mathcal{H} indicates an X.509 PKI-based authentication and \mathcal{L} indicates a general username password authentication). The hypothetic attack is shown in Fig. 3.

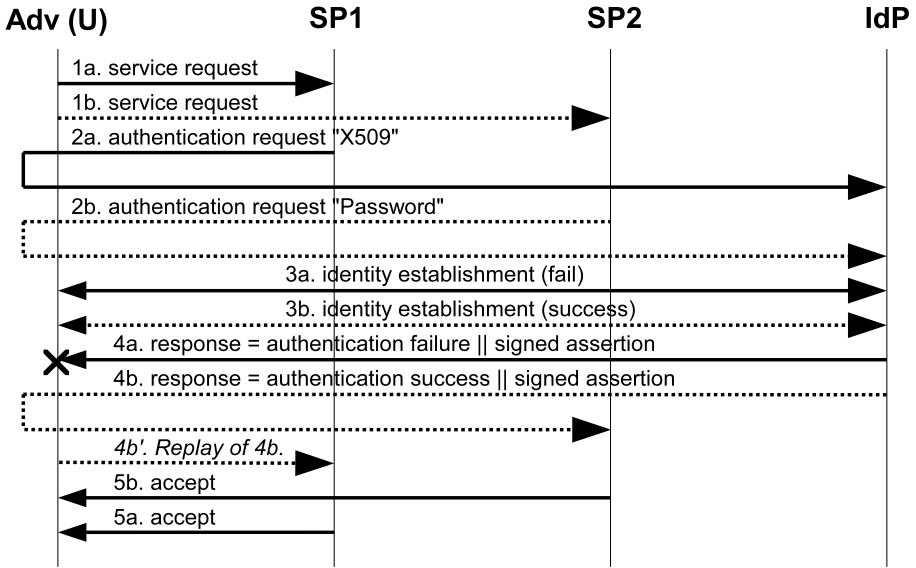


Fig. 3. Weakest Link Attack on Single Sign-On

3.2 Attack Damage

As we seen from the above illustration, the attack results in an acceptance of service request made by Adv who fails to satisfy the authentication context requirement of SP1 originally. If we consider Adv as an adversary who breaks only the victim user’s password, Adv in this case is as powerful as having forged an X.509 digital signature! In the other words, the weakest link attack enables an adversary to succeed in an authentication at a higher-level when she has only succeeded in breaking the one at a lower-level.

4 Weakest Link Attack on SAML V2.0 Web SSO

We now exploit our attack to a web SSO instantiated according to the specification in SAML V2.0 [3]. The instantiation only implements all mandatory procedures and elements as specified in the Standard, except the SPs specify the required authentication levels in `AuthnContextClassRef` in additional.

Remark 2. `AuthnContextClassRef` is contained in `<AuthnContext>` while the latter is contained in `<AuthnStatement>`. In SAML V2.0 web SSO, it is not mandatory for `<Assertion>` to include `<AuthnStatement>`.

4.1 Attack on SAML V2.0 Web SSO

An attack on an instantiation of SAML V2.0 web SSO is depicted in Fig. 4. The attack settings including the adversary power are the same as those described

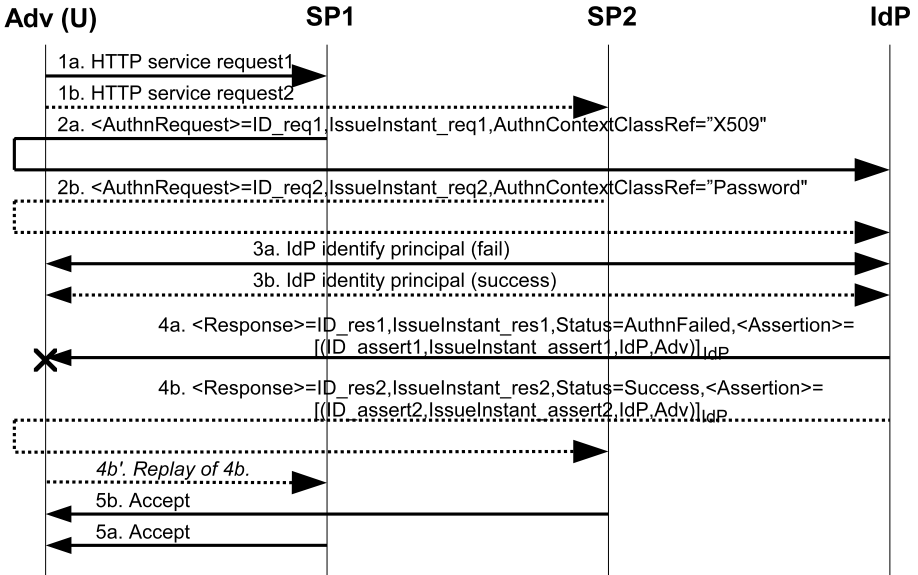


Fig. 4. Weakest Link Attack on SAML V2.0 web SSO

in Section 3. As per the authentication context requirements, SP1 specifies "X509" in `AuthnContextClassRef` while SP2 specifies "Password" in `AuthnContextClassRef`, which indicates a requirement of X.509 challenge and response authentication and a user name password authentication respectively.

As seen from Fig. 4, Adv succeeds in both service requests at SP1 and SP2 in the end. SP1 accepts the replay `<Response>` since `IssueInstant_res2` \approx `IssueInstant_res1` which is what she expects. This is because the two responses were generated in parallel. Moreover, `ID_res2` is unique to her (unless she verifies with SP2 in real time, but this is not a scalable practice).

Remark 3. Our attack result can also be achieved by modifying the `AuthnContextClassRef` attribute in the `<AuthnRequest>` message, but this can be prevented by message integrity mechanisms such as digital signatures.

Remark 4. Our attack as described above can be *prevented* by specifying the `Destination XML` attribute in the root element of a signed message. But we shall provide a repair in a generalized sense.

5 Results and Attack Generalization

The weakest link attack is successfully exploited in a default instantiation of SAML V2.0 web SSO. We further generalize our attack into the theorem below:

Theorem 1 (Generalization of Weakest Link Attack). *In a generic SSO system where underlying SSL/TLS is in place, the weakest link attack enables*

an adversary to succeed at all authentication levels associate to a victim user by breaking only the weakest authentication context. The attack is attainable if the following conditions are met in the corresponding SSO protocol:

1. *User redirection for request and response messages is required.*
2. *At least one of the following conditions in the response message is true:*
 - (a) *The level of authentication is not indicated or implied.*
 - (b) *The subject being authenticated is not indicated or implied.*
 - (c) *There is no integrity for the message segment that binds the subject and the authentication level.*

Proof. (A sketch.) Adv runs parallel authentication requests at two SPs and obtains `<response1>` and `<response2>`. User redirection enables Adv to intercept the unfavorable response, then replay the favorable one. Absence of authentication level indicator makes the receiving SP to accept the response inevitably. Absence of subject indicator enables Adv to replay responses correspond to other users. Alternately, a presence of authentication level indicator or subject indicator in a response without message integrity allows Adv to modify such values.

6 Repair

We propose the repair to the generalized weakest link attack below:

Corollary 1 (Repair to Weakest Link Attack). *The weakest link attack can be prevented when all of the followings are mandatory:*

1. *Include the indicator of the required authentication level in the response.*
2. *Include the indicator of the authentication subject in the response.*
3. *Sign the response message.*

Repair for SAML V2.0 web SSO. We suggest the following modifications:

1. Make the `AuthnContextClassRef` attribute as well as its container `<AuthnContext>` mandatory (i.e. [Required]) in the `<Assertion>`.
2. The entire `<Response>` element MUST be signed.

We also recommend the usage of the `InResponseTo` attribute, which is a reference to the identifier of the request to which the response corresponds. Upon receiving `<Response>`, SP SHOULD verify that the value of `InResponseTo` matches that of `ID_req` she generated in the corresponding `<AuthnRequest>`. A suggested `<Response>` element is given in the appendix.

7 Conclusion

In the above sections, we have presented a new parallel session attack specific to SSO. We started by an attack hypothesized in a generic SSO scenario. And then we evaluated the attack hypothesis by exploiting it on SAML V2.0 web SSO and

we succeeded. We generalized the attack into a theorem and proposed the corresponding repair. Due to the scope of SAML, we do not regard we have broken the SSO in SAML and its related specifications. Nevertheless we achieved significant attack results and contributed valuable information and considerations to SSO practitioners.

References

1. Thomas Groß. Security analysis of the saml single sign-on browser/artifact profile. In *Proceedings of the 19th Annual Computer Security Applications Conference*, December 2003.
2. OASIS SSTC. *Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML)*, November 2002.
3. OASIS SSTC. *Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0*, March 2005.
4. OASIS SSTC. *Bindings for the OASIS Security Assertion Markup Language (SAML) V2.0*, March 2005.
5. OASIS SSTC. *Profiles for the OASIS Security Assertion Markup Language (SAML) V2.0*, March 2005.
6. OASIS SSTC. *Security and Privacy Considerations for the OASIS Security Assertion Markup Language (SAML) V2.0*, March 2005.
7. OASIS SSTC. *SSTC Response to "Security Analysis of the SAML Single Sign-on Browser/Artifact Profile"*, July 2005.
8. Birgit Pfitzmann and Michael Waidner. Analysis of liberty single-signon with enabled clients. *IEEE Internet Computing*, 7(6):38–44, November 2003.
9. Jakob Skriver Steffen M. Hansen and Hanne Riis Nielson. Using static analysis to validate the saml single sign-on protocol. In *Proceedings of the 2005 Workshop on Issues in the Theory of Security*, pages 27–40, January 2005.

A Example of the Suggested <Response> Element

```

<Response
  ID="_b7042887-ad61-4fce-8b98-e2927324b986"
  IssueInstant="2005-10-14T00:32:01Z" Version="2.0"
  InResponseTo="_c4056889-de82-5dca-7c81-a342f324bcb0"
  xmlns="urn:oasis:names:tc:SAML:2.0:protocol"
  xmlns:saml="urn:oasis:names:tc:SAML:2.0:assertion">
  <saml:Issuer>https://www.anonymous.org/IDP</saml:Issuer>
    <ds:Signature xmlns:ds="http://www.w3.org/2000/09/xmldsig#">
      <ds:SignedInfo>
        <ds:CanonicalizationMethod
          Algorithm="http://www.w3.org/2001/10/xml-exc-c14n#" />
        <ds:SignatureMethod
          Algorithm="http://www.w3.org/2000/09/xmldsig#rsa-sha1" />
        <ds:Reference URI="#_c7055387-af61-4fce-8b98-e2927324b306">
          <ds:DigestMethod
            Algorithm="http://www.w3.org/2000/09/xmldsig#sha1" />
          <ds:DigestValue>TCDVsuG6grhyHbzhQFWFzGrxIPE=</ds:DigestValue>
        </ds:Reference>
      </ds:SignedInfo>
    <ds:SignatureValue>
      x/GyPbzmfEe85pGD3claXG4Vspb9V9jGCjwCRCKrtwPS6vdVNCcY5rHaFPYWkf+5
      EIYcPzx+pX1h43smwviCqXRjRtMANWbHlHWAptaKlywS7gFgsD01qjyen3CP+m3D
      w6vKhaqledl0BYyrIzb4KkHO4ahNyBVXbJwqv5pUae4=
    </ds:SignatureValue>
    <ds:KeyInfo>
      <ds:X509Data>
        <ds:X509Certificate>
          bmZvcmlhdGlvbiBUZWNobm9sb2d5MSUwIiwYDVQDEExIRVBLSBTZXJ2ZXIgdQ0Eg
          LS0gMjAwMjA3MDFBMB4XDTAyMDcyNjA3Mjc1MVoXDTA2MDkwNDA3Mjc1MVowGySx
          CzAJBgNVBAYTA1VTMREwDwYDVQQIEWhNaWNoaWdhbjESMBAGAlUEBxMJQW5uIEFy
          dENMCUGCSqGSIb3DQEJARYYcm9vdEBzAGlms5pbmRlcm5ldDIuZWR1MIGfMA0G
          pmqOI fGTWQIDAQABox0wGzAMBgnVHRMBAf8EAjAAMAsGA1UdDwQEAwIFoDANBgkq
          hkiG9w0BAQQFAAOBQBfDqEW+OI3jqBQHIBzhujN/PizdN7s/z4D5d3pptWDJf2n
          qgi71FV6MDkhmTvTqBtjmk3No7v/dnP6Hr7wHxvCCRwubnmIFZ6QZAv2FU78pLX
          8I3bsbmRAUg4UP9hH6ABVq4KQKMknxulxQxLhpR1ylGPdiowMNTREg8cCx3w/w==
        </ds:X509Certificate>
      </ds:X509Data>
    </ds:KeyInfo>
  </ds:Signature>
  <Status>
    <StatusCode Value="urn:oasis:names:tc:SAML:2.0:status:Success" />
  </Status>
  <Assertion ID="_a75adf55-01d7-40cc-929f-dbd8372ebdfc"
    IssueInstant="2005-10-15T00:24:02Z" Version="2.0"
    xmlns="urn:oasis:names:tc:SAML:2.0:assertion">
    <Issuer>https://www.anonymous.org/IDP</Issuer>
    <Subject>
      <NameID
        Format="urn:oasis:names:tc:SAML:1.1:nameid-format:emailAddress">
        anon@example.org
      </NameID>
      <SubjectConfirmation
        Method="urn:oasis:names:tc:SAML:2.0:cm:bearer" />
    </Subject>
    <AuthnStatement AuthnInstant="2005-10-15T00:24:02Z">
      <AuthnContext>
        <AuthnContextClassRef
          urn:oasis:names:tc:SAML:2.0:ac:classes:Password
        </AuthnContextClassRef>
      </AuthnContext>
    </AuthnStatement>
  </Assertion>
</Response>

```

Fig. 5. Example of the Suggested <Response> Element

An Inter-domain Key Agreement Protocol Using Weak Passwords*

Youngsook Lee, Junghyun Nam, and Dongho Won**

Information Security Group, Sungkyunkwan University, Korea
{yslee, jhnam, dhwon}@security.re.kr

Abstract. There have been many protocols proposed over the years for password authenticated key exchange in the three-party scenario, in which two clients attempt to establish a secret key interacting with one same authentication server. However, little has been done for password authenticated key exchange in the more general and realistic four-party setting, where two clients trying to establish a secret key are registered with different authentication servers. In fact, the recent protocol by Yeh and Sun seems to be the only password authenticated key exchange protocol in the four-party setting. But, the Yeh-Sun protocol adopts the so called “hybrid model”, in which each client needs not only to remember a password shared with the server but also to store and manage the server’s public key. In some sense, this hybrid approach obviates the reason for considering password authenticated protocols in the first place; it is difficult for humans to securely manage long cryptographic keys. In this paper, we propose a new protocol designed carefully for four-party password authenticated key exchange that requires each client only to remember a password shared with its authentication server. To the best of our knowledge, our new protocol is the first password-only authenticated key exchange protocol in the four-party setting.

Keywords: Key exchange, authentication, password, public key.

1 Introduction

Key exchange protocols are cryptographic protocols enabling two or more parties communicating over a public network to establish a common secret key that should be known only to the parties at the end of a protocol execution. This secret key, commonly called a *session key*, is then typically used to build confidential or integrity-protected communication channel between the parties. This means that key exchange should be linked to authentication so that each party has assurance that a session key is in fact shared with the intended party, and not an imposter. Hence, authenticated key exchange is of fundamental importance to anyone interested in communicating securely over a public network; even if

* This work was supported by the Korean Ministry of Information and Communication under the Information Technology Research Center (ITRC) support program supervised by the Institute of Information Technology Assessment (IITA).

** Corresponding author.

it is computationally infeasible to break the cryptographic algorithm used, the whole system becomes vulnerable to all manner of attacks if the keys are not securely established.

Achieving any form of authentication in key exchange protocols inevitably requires some secret information to be established between the communicating parties in advance of the authentication stage. Cryptographic keys, either secret keys for symmetric cryptography or private/public keys for asymmetric cryptography, may be one form of the underlying secret information pre-established between the parties. However, these high-entropy cryptographic keys are random in appearance and thus are difficult for humans to remember, entailing a significant amount of administrative work and costs. Eventually, it is this drawback that password-based authentication came to be widely used in reality. Passwords are drawn from a relatively small spaces like a dictionary and are easier for humans to remember than cryptographic keys with high entropy.

Bellovin and Merritt [5] was the first to consider how two parties, who only share a weak, low-entropy password, and who are communicating over a public network, authenticate each other and agree on a high-entropy cryptographic key to be used for protecting their subsequent communication. Their protocol, known as encrypted key exchange, or EKE, was a great success in showing how one can exchange password authenticated information while protecting poorly-chosen passwords from the notorious *off-line dictionary attacks*. Due in large part to the practical significance of password-based authentication, this initial work has been followed by a number of two-party protocols (e.g., [6, 3, 14, 29, 13]) offering various levels of security and complexity.

While two-party protocols for password authenticated key exchange (PAKE) are well suited for client-server architectures, they are inconvenient and costly for use in large scale peer-to-peer systems. Since two-party PAKE protocols require each potential pair of communication users to share a password, a large number of users results in an even larger number of potential passwords to be shared. It is due to this problem that three-party models have been often considered in designing PAKE protocols [12, 24, 17, 20, 1]. In a typical three-party setting, users (called clients) do not need to remember and manage multiple passwords, one for each communicating party; rather, each client shares a single password with a trusted server who then assists two clients in establishing a session key by providing authentication services to them. However, this convenience comes at the price of users' complete trust in the server. Therefore, whilst the three-party model will not replace the two-party model, it offers easier alternative solutions to the problem of password authenticated key exchange in peer-to-peer network environments.

One option for designing three-party PAKE protocols is to use a *hybrid* model in which each client stores at least one public key of its authentication server in addition to sharing a password with the server [22, 20, 28, 25, 7, 26]. It is probably fair to say that the hybrid model is mostly used because it provides an easy way to prevent both *off-line password guessing attack* and *undetectable on-line password guessing attack* [9]. However, the hybrid setting suffers from the disad-

vantage that clients must store a public key for each server to whom they wish to authenticate. If a client will need to deal with multiple servers, the client must store multiple public keys. In some sense, this obviates the reason for considering password authenticated protocols in the first place; human users cannot remember and is difficult to securely manage long cryptographic keys. This drawback has motivated research on password-only protocols [18, 19, 21] in which clients need to remember and manage only a short password shared with the server.

Up to now, most of literature discussing the problem of password authenticated key exchange focused their attention to the two-party model or the three-party model. While the three-party model seems to provide a more realistic approach in practice than the two-party model in which clients are expected to store multiple passwords, it is still restrictive in the sense that it assumes that two clients are registered and authenticated by the same server. In reality, the authentication server of one client may be different from that of another client. Indeed, it is a typical environmental assumption [16] that a client registers with the server in its own realm and trusts only its own server, not the servers in other realms. In this case, how to efficiently authenticate a client who is registered with the server in the other realm becomes an important issue.

This paper considers password authenticated key exchange in the inter-domain distributed computing environment just mentioned above. We propose a new inter-domain PAKE protocol which involves four participants: two clients and two authentication servers, one for each client. To the best of our knowledge, the recent protocol by Yeh and Sun [27] is the only PAKE protocol in the four-party model. However, the Yeh-Sun protocol employs the hybrid model and so each client in the protocol needs to manage the server's public key in addition to remembering a password shared with the server. In fact, our construction seems to be the first four-party PAKE protocol that requires each client only to share a password with its authentication server.

2 Protocol Preliminaries

We begin with the requisite definitions. There are four entities involved in the protocol: two clients A and B , and two authentication servers SA and SB respectively of A and B . We denote by ID_A , ID_B , ID_{SA} , and ID_{SB} , the identities of A , B , SA , and SB , respectively.

Computational Diffie-Hellman (CDH) Assumption. Let g be a fixed generator of the finite cyclic group \mathbb{Z}_p^* . Informally, the CDH problem is to compute g^{ab} given g^a and g^b , where a and b were drawn at random from $\{1, \dots, |\mathbb{Z}_p^*|\}$. Roughly stated, \mathbb{Z}_p^* is said to satisfy the CDH assumption if solving the CDH problem in \mathbb{Z}_p^* is computationally infeasible for all probabilistic polynomial time algorithms.

Symmetric Encryption Scheme. A symmetric encryption scheme is a triple of polynomial time algorithms $\Gamma = (\mathcal{K}, \mathcal{E}, \mathcal{D})$ such that:

- The *key generation algorithm* \mathcal{K} is a randomized algorithm that returns a key k . Let $\text{Keys}(\Gamma)$ be the set of all keys that have non-zero probability of being output of \mathcal{K} .
- The *encryption algorithm* \mathcal{E} takes as input a key $k \in \text{Keys}(\Gamma)$ and a plaintext $m \in \{0, 1\}^*$. It returns a ciphertext $\mathcal{E}_k(m)$ of m under the key k . This algorithm might be randomized or stateful.
- The *deterministic decryption algorithm* \mathcal{D} takes as input a key $k \in \text{Keys}(\Gamma)$ and a purported ciphertext $c \in \{0, 1\}^*$. It returns $\mathcal{D}_k(c)$, which is a plaintext $m \in \{0, 1\}^*$ or a distinguished symbol \perp . The return value \perp indicates that the given ciphertext c is invalid for the key k .

We say, informally, that a symmetric encryption scheme Γ is secure if it ensures confidentiality of messages under chosen-ciphertext attack (CCA) and guarantees integrity of ciphertexts [23]. As shown in [2, 15], this combination of security properties implies indistinguishability under CCA which, in turn, is equivalent to non-malleability [10] under CCA.

Signature Scheme. A digital signature scheme is a triple of algorithms $\Sigma = (\mathcal{G}, \mathcal{S}, \mathcal{V})$ such that:

- The *probabilistic key generation algorithm* \mathcal{G} , on input a security parameter 1^ℓ , outputs a pair of matching public and private keys (PK, SK) .
- The *signing algorithm* \mathcal{S} is a probabilistic polynomial time algorithm that, given as input a message m and a key pair (PK, SK) , outputs a signature σ of m .
- The *verification algorithm* \mathcal{V} is a polynomial time algorithm that on input (m, σ, PK) , outputs 1 if σ is a valid signature of the message m with respect to PK , and 0 otherwise.

We say that a signature scheme Σ is secure if the probability of succeeding with an existential forgery under adaptive chosen message attack [11] is negligible for all probabilistic polynomial time attackers.

Initialization. During some initialization phase, two servers SA and SB agree on the following public parameters: a large prime p , a generator g of \mathbb{Z}_p^* satisfying the CDH assumption, a one-way hash function H , a secure symmetric encryption scheme $\Gamma = (\mathcal{K}, \mathcal{E}, \mathcal{D})$, and a secure signature scheme $\Sigma = (\mathcal{G}, \mathcal{S}, \mathcal{V})$. In addition, public/private key pairs are generated for each server by running the key generation algorithm $\mathcal{G}(1^\ell)$. We denote by (PK_X, SK_X) the public/private keys of the server X for $X \in \{SA, SB\}$. As part of the initialization, the client A (resp. B) registers with server SA (resp. SB), by choosing PW_A (resp. PW_B) and sending it to the server via a secure channel.

3 A Four-Party Password Authenticated Key Agreement Protocol

In this section we present a new four-party password authenticated key agreement protocol in which two clients wishing to agree on a session key do not need

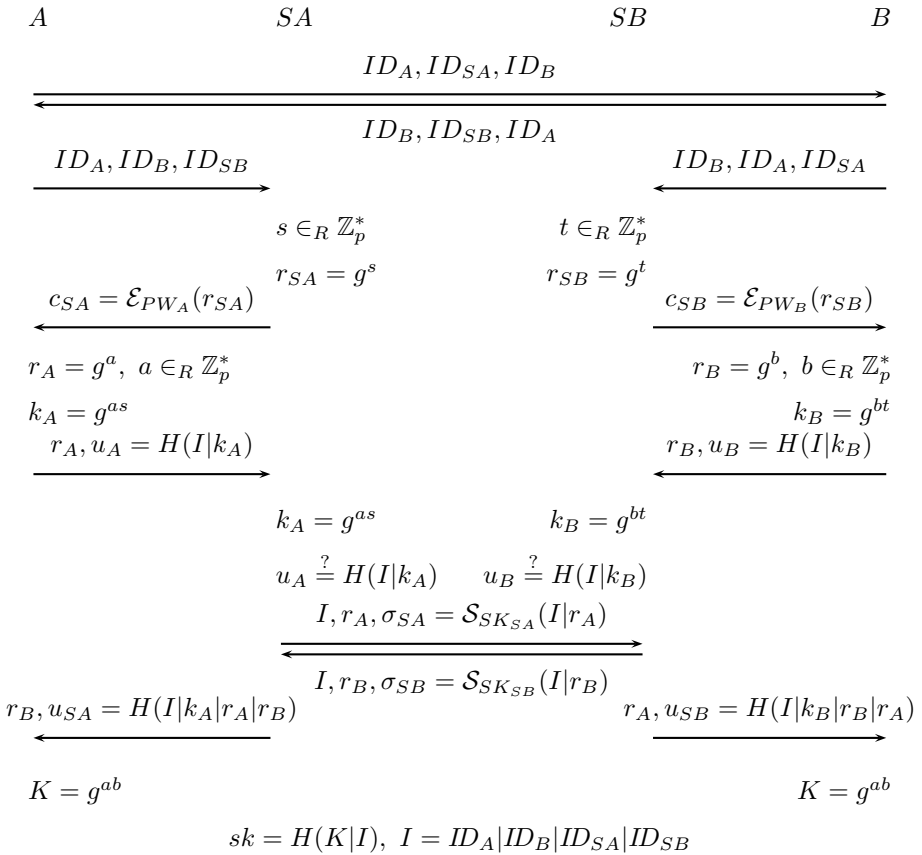


Fig. 1. Four-party password authenticated key agreement protocol

to store any public key of their authentication server but only need to share a short, easy-to-remember password with the server. In describing the protocol, we will omit ‘mod p ’ from expressions for notational simplicity. The proposed protocol is outlined in Fig. 1 and a more detailed description is as follows:

1. Two clients A and B first need to inform each other of their respective authentication server. To this end, A sends to B the message $\langle ID_A, ID_{SA}, ID_B \rangle$ and B sends to A the message $\langle ID_B, ID_{SB}, ID_A \rangle$.
2. The clients request the assistance of their respective server in establishing a session key between them. Client A (resp. B) does this by sending the message $\langle ID_A, ID_B, ID_{SB} \rangle$ (resp. $\langle ID_B, ID_A, ID_{SA} \rangle$) to the server SA (resp. SB).
3. Server SA chooses a random number $s \in_R \mathbb{Z}_p^*$, computes $r_{SA} = g^s$ and $c_{SA} = \mathcal{E}_{PW_A}(r_{SA})$, and sends the ciphertext c_{SA} to A . Similarly, server SB chooses a random number $t \in_R \mathbb{Z}_p^*$, computes $r_{SB} = g^t$ and $c_{SB} = \mathcal{E}_{PW_B}(r_{SB})$, and sends the ciphertext c_{SB} to B .

4. After receiving c_{SA} , client A recovers r_{SA} by decrypting c_{SA} , i.e., $r_{SA} = \mathcal{D}_{PW_A}(c_{SA})$, and chooses a random number $a \in_R \mathbb{Z}_p^*$. Given r_{SA} and a , client A computes the one time key k_A to be shared with the server SA as

$$k_A = g^{as} = (r_{SA})^a.$$

Additionally, A computes $r_A = g^a$ and $u_A = H(ID_A|ID_B|ID_{SA}|ID_{SB}|k_A)$. Then A sends the message $\langle r_A, u_A \rangle$ to the server SA .

Meanwhile, client B , having received c_{SB} , computes $r_{SB} = \mathcal{D}_{PW_B}(c_{SB})$ and chooses a random number $b \in_R \mathbb{Z}_p^*$. From r_{SB} and b , B computes the one time key k_B to be shared with SB as

$$k_B = g^{bt} = (r_{SB})^b.$$

B also computes $r_B = g^b$ and $u_B = H(ID_A|ID_B|ID_{SA}|ID_{SB}|k_B)$ and then sends $\langle r_B, u_B \rangle$ to SB .

5. Upon receiving $\langle r_A, u_A \rangle$, server SA first computes the one time key $k_A = g^{as}$ shared with A and then verifies that u_A from A equals the hash value $H(ID_A|ID_B|ID_{SA}|ID_{SB}|k_A)$. If the verification fails, SA stops executing the protocol; otherwise, SA believes that client A is genuine. SA then sends the message $\langle ID_A, ID_B, ID_{SA}, ID_{SB}, r_A, \sigma_{SA} \rangle$ to the server SB , where σ_{SA} is the signature on $ID_A|ID_B|ID_{SA}|ID_{SB}|r_A$ generated by the signing algorithm \mathcal{S} using the private key SK_{SA} , namely,

$$\sigma_{SA} = \mathcal{S}_{SK_{SA}}(ID_A|ID_B|ID_{SA}|ID_{SB}|r_A).$$

Similarly, upon receiving $\langle r_B, u_B \rangle$, server SB computes the one time key $k_B = g^{bt}$ shared with B and verifies that u_B from B equals $H(ID_A|ID_B|ID_{SA}|ID_{SB}|k_B)$. If the verification fails, SB aborts the protocol; otherwise, SB believes client B as authentic. Then SB sends the message $\langle ID_A, ID_B, ID_{SA}, ID_{SB}, r_B, \sigma_{SB} \rangle$ to SA , where σ_{SB} is the signature on $ID_A|ID_B|ID_{SA}|ID_{SB}|r_B$ generated by \mathcal{S} using the private key SK_{SB} , namely,

$$\sigma_{SB} = \mathcal{S}_{SK_{SB}}(ID_A|ID_B|ID_{SA}|ID_{SB}|r_B).$$

6. After receiving $\langle ID_A, ID_B, ID_{SA}, ID_{SB}, r_B, \sigma_{SB} \rangle$, SA first verifies the signature σ_{SB} using the public key PK_{SB} . SA halts immediately if the verification fails. Otherwise, SA computes

$$u_{SA} = H(ID_A|ID_B|ID_{SA}|ID_{SB}|k_A|r_A|r_B)$$

and sends $\langle r_B, u_{SA} \rangle$ to client A .

The server SB , upon receiving $\langle ID_A, ID_B, ID_{SA}, ID_{SB}, r_A, \sigma_{SA} \rangle$, verifies the signature σ_{SA} , and if correct, computes

$$u_{SB} = H(ID_A|ID_B|ID_{SA}|ID_{SB}|k_B|r_B|r_A)$$

and sends $\langle r_A, u_{SB} \rangle$ to B .

7. The client A checks whether u_{SA} from SA equals $H(ID_A|ID_B|ID_{SA}|ID_{SB}|k_A|r_A|r_B)$. If this is untrue, A aborts the protocol. Otherwise, A computes the common secret value K as

$$K = g^{ab} = r_B^a.$$

Similarly, client B verifies that u_{SB} from SB equals $H(ID_A|ID_B|ID_{SA}|ID_{SB}|k_B|r_B|r_A)$, and if the verification succeeds, computes the common secret value K as

$$K = g^{ab} = r_A^b.$$

Finally, the clients compute their session key sk as

$$sk = H(K|ID_A|ID_B|ID_{SA}|ID_{SB}).$$

4 Security Analysis

In this preliminary version of the paper, we only provide a heuristic security analysis of the proposed protocol, considering a variety of attacks and security properties; a rigorous proof of security in a formal communication model will be given in the full version of this paper.

Off-Line Password Guessing Attack. In this attack, an attacker may try to guess a password and then to check the correctness of the guessed password off-line. If his guess fails, the attacker tries again with another password, until he find the proper one. In the proposed protocol, the only information related to passwords is $c_{SA} = \mathcal{E}_{PW_A}(g^s)$ and $c_{SB} = \mathcal{E}_{PW_B}(g^t)$, but because s and t are chosen randomly, these values does not help the attacker to verify directly the correctness of the guessed passwords. Thus, off-line password guessing attacks would be unsuccessful against the proposed protocol.

Undetectable On-Line Password Guessing Attack. At the highest level of security threat to password authenticated key exchange protocols are undetectable on-line password guessing attacks [9] where an attacker tries to check the correctness of a guessed password in an on-line transaction with the server, i.e., in a fake execution of the protocol; if his guess fails, he starts a new transaction with the server using another guessed password. Indeed, the possibility of an undetectable on-line password guessing attack in the three-or-more-party setting represents a qualitative difference from the two-party setting where such attack is not a concern. However, this attack is meaningful only when the server is unable to distinguish an honest request from a malicious one, since a failed guess should not be detected and logged by the server.

In our protocol, the server is the first who issues a challenge and the client is the first who replies with an answer to some challenge. It is mainly due to this ordering that the protocol is secure against undetectable on-line password guessing attacks. Suppose that an attacker A' , posing as A , decrypts c_{SA} by guessing

a password, computes r_A , k_A , and $u_A = H(ID_A|ID_B|ID_{SA}|ID_{SB}|k_A)$ by choosing his own random a' , and sends the fake message $\langle r_A, u_A \rangle$ to server SA . Then, the server SA , upon receiving r_A and u_A from A' , should be easily able to detect a failed guess since the protocol specification mandates SA to check the correctness of u_A . Note that the attacker cannot send a correct u_A without knowing a correct k_A which in turn only can be computed if the guessed password is correct. Hence, the proposed protocol can resist undetectable on-line password guessing attacks.

Insider Attack. One of the main differences between the two-party setting and the three-or-more-party setting is the existence of insider attacks, i.e., attacks by malicious insiders. Namely, insider attacks are a concern specific to the three-or-more-party setting, and do not need to be considered in the two-party setting.

Suppose one client, say A , tries to learn the password of the other client B during execution of the protocol. Of course, A should not be able to have this ability through a protocol run and this is still an important security concern to be addressed. As mentioned earlier in this section, the only information related to the B 's password is $c_{SB} = \mathcal{E}_{PW_B}(r_{SB})$, where r_{SB} is computed as $r_{SB} = g^t$ and t is a random value chosen by the server SB . The actual value r_{SB} itself is never included in any message sent in the protocol and t is a secret information only known to SB . This means that the malicious insider A has no advantage over outside attackers in learning B 's password. In other words, A 's privileged information — PW_A , r_{SA} , and k_A — gives A no help in learning B 's password. Therefore, our protocol is also secure against insider attacks.

Implicit Key Authentication. The fundamental security goal for a key exchange protocol to achieve is implicit key authentication. Loosely stated, a key exchange protocol is said to achieve implicit key authentication if each party trying to establish a session key is assured that no other party aside from the intended parties can learn any information about the session key. Here, we restrict our attention to passive attackers; active attackers will be considered in the full version of this paper.

Given $r_A = g^a$ and $r_B = g^b$, the secret value $K = g^{ab}$ cannot be computed, since no polynomial algorithm has been found to solve the computational Diffie-Hellman problem. Thus, if the random numbers a and b are unknown, then the session key sk cannot be computed since H is a one-way hash function. Hence, the secrecy of the session key is guaranteed based on the computational Diffie-Hellman assumption in the random oracle model [4].

Perfect Forward Secrecy. Perfect forward secrecy [8] is provided if past session keys are not recovered by compromise of some underlying long-term information at the present time. The proposed protocol also achieves perfect forward secrecy since the long-term information of the protocol participants is used for implicit key authentication only, and not for hiding the session key. Even if an attacker obtains the clients' passwords and the servers' signing private keys, he

cannot get any of secret values $K = g^{ab}$ computed in previous sessions since a and b chosen respectively by A and B are unknown. Because the session key in this protocol is computed as $sk = H(K|ID_A|ID_B|ID_{SA}|ID_{SB})$, he cannot compute it without knowing the secret value K . Hence, our protocol does indeed achieve perfect forward secrecy.

Known Key Security. Known key security is said to be provided if compromise of some session keys does not help an attacker learn about any other session keys or impersonate a party in some later session. In our protocol, the session keys generated in different sessions are independent since the ephemeral secret exponents a and b are chosen independently at random from session to session. Thus, the proposed protocol still achieves its goal in the face of an attacker who has learned some other session keys.

References

1. M. Abdalla, P.-A. Fouque, and D. Pointcheval, Password-based authenticated key exchange in the three-party setting, *PKC 2005*, LNCS 3386, pp. 65–84, 2005.
2. M. Bellare and C. Namprempe, Authenticated encryption: Relations among notions and analysis of the generic composition paradigm, *Asiacrypt 2000*, LNCS 1976, pp. 531–545, 2000.
3. M. Bellare, D. Pointcheval, and P. Rogaway, Authenticated key exchange secure against dictionary attacks, *Eurocrypt 2000*, LNCS 1807, pp. 139–155, 2000.
4. M. Bellare and P. Rogaway, Random oracles are practical: A paradigm for designing efficient protocols, *Proc. ACM CCS 1993*, pp. 62–73, 1993.
5. S.M. Bellovin and M. Merritt, Encrypted key exchange: password-based protocols secure against dictionary attacks, *Proc. IEEE Symp. Research in Security and Privacy*, pp. 72–84, 1992.
6. V. Boyko, P. MacKenzie, and S. Patel, Provably secure password-authenticated key exchange using Diffie-Hellman, *Eurocrypt 2000*, LNCS 1807, pp. 156–171, 2000.
7. H.-Y. Chien and J.-K. Jan, A hybrid authentication protocol for large mobile network, *The Journal of Systems and Software*, vol. 67, no. 2, pp. 123–130, 2003.
8. W. Diffie, P. Oorschot, and M. Wiener, Authentication and authenticated key exchanges, *Designs, Codes, and Cryptography*, vol. 2, no. 2, pp. 107–125, 1992.
9. Y. Ding and P. Horster, Undetectable on-line password guessing attacks, *ACM SIGOPS Operating Systems Review*, vol. 29, no. 4, pp. 77–86, 1995.
10. D. Dolev, C. Dwork, and M. Naor, Nonmalleable cryptography, *SIAM Journal on Computing*, vol. 30, no. 2, pp. 391–437, 2000.
11. S. Goldwasser, S. Micali, and R. Rivest, A digital signature scheme secure against adaptive chosen-message attacks, *SIAM Journal of Computing*, vol. 17, no. 2, pp. 281–308, 1988.
12. L. Gong, M.-L. Lomas, R.-M. Needham, and J.-H. Saltzer, Protecting poorly chosen secrets from guessing attacks, *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 5, pp. 648–656, 1993.
13. S. Jiang and G. Gong, Password based key exchange with mutual authentication, *SAC 2004*, LNCS 3357, pp. 267–279, 2005.
14. J. Katz, R. Ostrovsky, and M. Yung, Efficient password-authenticated key exchange using human-memorable passwords, *Eurocrypt 2001*, LNCS 2045, pp. 475–494, 2001.

15. J. Katz and M. Yung, Unforgeable encryption and adaptively secure modes of operation, *FSE 2000*, LNCS 1978, pp. 284–299, 2000.
16. J.-T. Kohl and B.-C. Neumann, The Kerberos Network Authentication Service, Version 5 Revision 5, Project Athena, Massachusetts Institute of Technology, 1992.
17. T. Kwon, M. Kang, S. Jung, and J. Song, An improvement of the password-based authentication protocol (K1P) on security against replay attacks, *IEICE Transactions on Communications*, vol. E82-B, no. 7, pp. 991–997, 1999.
18. T.-F. Lee, T. Hwang, and C.-L. Lin, Enhanced three-party encrypted key exchange without server public keys, *Computer & Security*, vol. 23, no. 7, pp. 571–577, 2004.
19. S.-W. Lee, H.-S. Kim, and K.-Y. Yoo, Efficient verifier-based key agreement protocol for three parties without server's public key, *Applied Mathematics and Computation*, vol. 167, no. 2, pp. 996–1003, 2005.
20. C.-L. Lin, H.-M. Sun, and T. Hwang, Three-party encrypted key exchange: attacks and a solution, *ACM SIGOPS Operating Systems Review*, vol. 34, no. 4, pp. 12–20, 2000.
21. C.-L. Lin, H.-M. Sun, M. Steiner, and T. Hwang, Three-party encrypted key exchange without server public-keys, *IEEE Communications letters*, vol. 5, no. 12, pp. 497–499, 2001.
22. T.-M.-A. Lomas, L. Gong, J.-H. Saltzer, and R.-M. Needham, Reducing risks from poorly chosen keys, *ACM SIGOPS Operating Systems Review*, vol. 23, no. 5, pp. 14–18, 1989.
23. P. Rogaway, M. Bellare, J. Black, and T. Krovetz, OCB: A block-cipher mode of operation for efficient authenticated encryption, *Proc. ACM CCS 2001*, pp. 196–205, 2001.
24. M. Steiner, G. Tsudik, and M. Waidner, Refinement and extension of encrypted key exchange, *ACM SIGOPS Operating Systems Review*, vol. 29, no. 3, pp. 22–30, 1995.
25. H.-M. Sun, B.-C. Chen, and T. Hwang, Secure key agreement protocols for three-party against guessing attacks, *The Journal of Systems and Software*, vol. 75, no. 1–2, pp. 63–68, 2005.
26. H.-T. Yeh and H.-M. Sun, Password-based user authentication and key distribution protocols for client-server applications, *The Journal of Systems and Software*, vol. 72, no. 1, pp. 97–103, 2004.
27. H.-T. Yeh, and H.-M. Sun, Password authenticated key exchange protocols among diverse network domains, *Computers and Electrical Engineering*, vol. 31, no. 3, pp. 175–189, 2005.
28. H.-T. Yeh, H.-M. Sun, and T. Hwang, Efficient three-party authentication and agreement protocols resistant to password guessing attacks, *The Journal of Information Science and Engineering*, vol. 19, pp. 1059–1070, 2003.
29. M. Zhang, New approaches to password authenticated key exchange based on RSA, *Asiacrypt 2004*, LNCS 3329, pp. 230–244, 2004.

A Practical Solution for Distribution Rights Protection in Multicast Environments

Josep Pegueroles, Marcel Fernández, Francisco Rico-Novella,
and Miguel Soriano

Telematics Engineering Department, Technical University of Catalonia,
Jordi Girona 1-3, CAMPUS NORD C3 08034, Barcelona, Spain
{josep, marcelf, telfrn, soriano}@entel.upc.edu

Abstract. One of the main problems that remain to be solved in pay-per-view Internet services is copyright protection. As in almost every scenario, any copyright protection scheme has to deal with two main aspects: protect the true content authors from those who may dishonestly claim ownership of intellectual property rights and prevent piracy by detecting the authorized (but dishonest) users responsible of illegal redistribution of copies. The former aspect can be solved with watermarking techniques while for the latter, fingerprinting mechanisms are the most appropriate ones. In internet services such as Web-TV or near video on-demand where multicast is used, watermarking can be directly applied. On the other hand, multicast fingerprinting has been seldom studied because delivering different marked content for different receivers seems a contradiction with multicast basics. In this paper we present a solution to prevent unauthorized redistribution of content in multicast scenarios. The system is based on a trusted soft-engine embedded in the receiver and co-managed by the content distributor. The trusted soft-engine is responsible of the client-side multicast key management functions. It only will allow the decryption and displaying of the actual data if it has previously inserted a fingerprinting mark with the identity of the decoder. Upon finding a pirate copy of any multicast delivered content, this mark can be used to unambiguously reveal the identity of the receiver that decoded the content from which the pirate copies are made.

1 Introduction

The use of multicast is an important handicap when adding fingerprinting mechanisms to commercial content delivery over the network. In multicast communications all the receivers get the same stream of bits. If anyone of the end users illegally redistributes the content, there is no way to distinguish the pirate copy delivered by the malicious user from the original ones hold by the honest users. So, unauthorized redistributors would be masked by the rest of the multicast group. In this sense, copyright protection for multicast arises as a challenging problem in commercial network applications.

Not many solutions have been previously presented addressing this problem. Moreover, all of them present many disadvantages: incompatibility with multicast encryption schemes or relying in the existence of trusted and programmable

intermediate network nodes. These disadvantages make the current proposals infeasible to be efficiently implemented in real scenarios. In this paper we present a practical, ready-to-implement solution to the multicast fingerprinting problem.

Our scheme is based on what we call the “trusted soft-engine” which is either a software or hardware tamper-proof entity embedded in the receiver at the client side. We emphasize that although this engine runs in the client side, some of their actions are actually managed by the server (distributor). In other words, the proposed solution is not only a client-side fingerprinting device and, of course, its server controlled part is assumed to be unbreachable. This engine has the following particularity with respect to the actions it performs: some actions require the management of the content distributor alone; other kinds of actions will require the management of the end user alone; finally, there are some actions that require the activity of a trusted third party.

The trusted soft-engine will be the responsible, on one hand, of receiving the stream of bits and decrypting it using the secure multicast group key. At the same time, it will embed a single mark to the actual data that is passed to the media player.

2 Required Background

2.1 Fingerprinting Schemes

Multicast networks allow the delivery of “digital goods” that are not even contained in a physical support. Once the buyer has received the content, he is free to store it in the physical support of his choice. Of course, he is also free to make as many (inexpensive) copies of the content as he wishes. This copying and storing easiness constitutes a severe threat to authorship and distribution rights.

Many experts share the opinion that e-commerce will not be a total reality, until this threat to intellectual property rights can be efficiently counteracted.

Among the mechanisms used in protecting intellectual property we observe two groups: copy prevention mechanisms and copy detection mechanisms. The failure of copy prevention schemes, such as the DVD copy prevention system, has shifted many research efforts towards the search of feasible copy detection mechanisms.

These efforts have brought the development of new proposals of copy detection systems for many different types of digital content.

A large number of these new proposals can be grouped under the name of watermarking. The watermarking technique consists of embedding a set of marks in each copy of the digital content. Since the embedded set of marks is the same for all copies, watermarking schemes ensure intellectual property protection but fail to protect distribution rights. If protection against distribution rights is also required then the concept of watermarking needs to be extended further off.

The functionality of watermarking can be extended by allowing the embedded set of marks to be different in each copy of the object. This true original idea was introduced by [17] and is called fingerprinting [2,11,17] because of its analogy to human fingerprints.

The copies of a digital object obtained under a fingerprinting scheme are, of course, all different. Having distinct copies of the same object univocally identifies the buyer of a copy and therefore allows tracing illegal plain redistribution.

Unfortunately, the seemingly advantageous extension from watermarking to fingerprinting introduces weaknesses to the resulting scheme. These weaknesses come in the form of a collusion attack.

In a collusion attack [3,4] a group of dishonest users get together and by comparing their copies they are able to detect the symbols where their copies differ. By changing some of these symbols they create a new pirate copy that tries to hide their identities.

Consequently, an adequate fingerprinting scheme must allow identification of dishonest buyers who took part in a collusion. Detection of such dishonest buyers can be achieved by using sets of marks with traceability properties, such as the ones presented in [3,4,9,19]

2.2 Multicast Security

Aside from protecting the copyright, probably the first security requirement in multicast communications is protecting the data from unauthorized eavesdroppers [7]. This is easily achieved by means of encryption. The addition of encryption algorithms in multicast communications has several restrictions: it must not add significant delay to data transmission or excessive packet loss due to time constraint violation will occur, moreover all the delivered data shall be encrypted using a session key shared by all the multicast group members [18].

The shared key provides group secrecy and source authentication. This key must be updated every time membership in the group changes, if so Forward and Backward Secrecy (FS and BS) are provided. FS means that the session key gives no meaningful information about future session keys, that is to say, no leaving member can obtain information about future group communication. BS means that a session key provides no meaningful information about past session keys or that no joining member can obtain information about past group communication [12]).

The above mentioned constraints force us to an on-line encryption of the multimedia content (just before sending it) and to use different keys for protecting the same file from eavesdroppers, even during the same session. In order to avoid extra delay in the process, symmetric encryption is generally used. The most successful proposals in multicast key encryption and key management schemes are presented in next sections.

3 Previous Work

3.1 State Of the Art in Fingerprinting Techniques for Multicast.

Two distinct approaches have been presented in the literature to achieve multicast fingerprinting:

- A priori generation of different copies of the same content with different watermarks and following the appropriate selection of packets during distribution. This leads to allow unique identifiers at the receiver end.
- Delivery of a single copy of unwatermarked content while middle entities in the distribution process are responsible of the introduction of the appropriate watermarks.

Next, we slightly overview the different proposals of these two approaches:

In [10], for a given multicast video, the sender applies two different watermark functions to generate two different watermarked frames for every frame in the stream. Each end user has a uniquely identifying random key. The length of this key is the number of video frames in the stream. This identifying key is also known to the distributor. For each watermarked frame in the stream a different key is used to encrypt it. The random bit stream determines whether a member will be given one key or another for decryption. If there is only one leaking member, the distributor can identify this dishonest redistributor because he can read the watermarks to produce the identifying key and because he knows the bitstreams of all members.

Another approach to the problem of copyright protection in multicast is Distributed watermarking (Watercasting) [5]. For a multicast routing tree with maximum depth (d), the distributor generates $n \geq d$ different copies of each packet. Each group of n alternate packets (each one with same content but different marks) is called a transmission group. On receiving a transmission group, a router forwards all but one of those packets to each downstream interface. Each last hop router in the distribution tree will receive $n - d_r$ packets from each transmission group, where d_r is the depth of the route to this router. Exactly one of these packets will be forwarded onto the subnet with receivers. This provides a stream for each receiver with a unique sequence of watermarked packets.

In [15], the authors introduce a hierarchy of intermediaries into the network. Each intermediary has a unique ID which is used to define the path from the source to the intermediary. The Path ID is embedded into the content to identify the path it has travelled. Each intermediary embeds its portion of the Path ID into the content before it forwards the content through the network. A watermark embedded by the distributor identifies the domain of a receiver. The last hop allows the intermediaries to mark the content uniquely for any child receivers. Multiple watermarks can be embedded using modified versions of existing algorithms.

In [8] it was presented the Hierarchical tagging that allows a content producer to insert a different watermark for each of his distributors. Similarly, each distributor can insert a watermark for several sub-distributors. This process can continue until the end-users receive tagged content identifying all the distributors and the producer. The content is hidden using cryptographic techniques. Each end user receives the same data, performs some processing and retrieves only the data prepared for him.

3.2 Current Proposals Drawbacks

All the above presented methods construct a different fingerprinted marked content to each end-user by either pre-processing techniques of the content or trusting the intermediate nodes of the network to perform some actions or both at the same time.

The first important drawback of these proposals is that they are not scalable and different copies of the data stream need to be watermarked and encrypted a priori. Moreover, they are not fully compatible with state of the art multicast encryption schemes. In many cases the pre-processing required by the distributor or the client creates a weakness either in terms of computational cost or security.

Another drawback when relying on intermediate network nodes is that the distributor needs the information about the tree topology if the illegal copy has to be

traced. In open networks such as Internet, this is usually difficult to know a priori. Moreover, network providers may not be willing to provide security functionality to intermediate routers.

3.3 State Of the Art in Multicast Encryption Techniques

With respect to the multicast encryption field, several works prevent new group members or leaving members from accessing data sent before they joined or after they leave. The simplest way a Key Server can deliver a new session key to the members is through a secret unicast connection with each of the remaining members of the group [13]. This solution presents the worst behaviour with respect to efficiency parameters. All figures have a dependency on the number of members in the multicast group.

In order to reduce the number of messages for rekeying in [1,6,13,16] Logical Key Tree Schemes were presented. Key tree schemes use two types of encryption keys: Session Encryption Keys (SEK) and Key Encryption Keys (KEK). SEKs are used to cipher the actual data that multicast groups exchange, for example, video streams in multicast near-on-demand video sessions. KEKs are used to cipher the keying material that members need in order to get the SEK. Normally, KEKs are structured in logical binary trees. All users know the root of the key tree and the leaf nodes are users' individual keys. Imagine, for instance, the Tree in Figure 1. Nodes will be referred as (level number, position at level), so we will refer to root node as (1,1); the sons of root node will be (2,1) and (2,2) and so on. Key in node (X,Y) will be noted as $K(X,Y)$.

Consider a group of 8 users. The tree has 15 nodes, each node corresponds to a KEK. Group members are located at leaf nodes. Keys in the leaves are only known by single users. $K(1,1)$ is known by all members in the group. The rest of the keys are revealed only to users considered sons of the node. For example, $K(3,1)$ is known only by users in leaves (4,1) and (4,2), and $K(2,2)$ is revealed only to nodes (4,5) to (4,8).

3.4 Logical Key Hierarchy (LKH)

The simplest logical tree key management scheme is LKH [14]. It works as follows. Consider the multicast group in Figure 1 with 8 members (M_1, \dots, M_7) and a Key Server (KS). Each member must store a subset of the controller's keys. This subset of KEKs will allow the member to get the new SEK when it changes. A generic member (M_j) stores the subset of keys in the path from the leaf where he is to the root. In our example, member M_1 , in node (4,1), will store $K(4,1), K(3,1), K(2,1)$ and $K(1,1)$.

When a member (M_8) leaves the group, the Key Server must update every KEK in the path from the leaf, where new member was located, to the root. See Figure 2 in which new keys are denoted with quotes.

The KS has to reveal the updated keys to the corresponding users. He uses the existing key hierarchy, along with reliable multicast, to efficiently distribute them as follows. He sends one message containing the set of updated keys to the members in node (4,7). This operation can be done via a unicast channel and using M_7 's individual key. After that, the key server constructs and sends a multicast message containing the updated keys ($K'(2,2)$, $K'(1,1)$) ciphered with $K(3,3)$, so only members in nodes (4,5) and (4,6) can decipher it. Finally, KS also constructs and

sends a multicast message containing the new root key ($K'(1,1)$) and ciphered with $K(2,1)$, so members in nodes (4,1) to (4,4) can decipher it. At this point, the remaining members in the multicast group know the subset of keys from their leaves to the root. Every member knows the root key, so this can be used to cipher a multicast message containing the new session key (SEK').

Following the example it is easy to see how LKH can update keys more efficiently than using only unicast connections with every end-user.

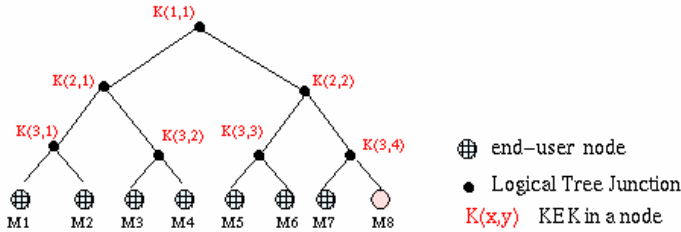


Fig. 1. Logical Key Hierarchy for multicast key management

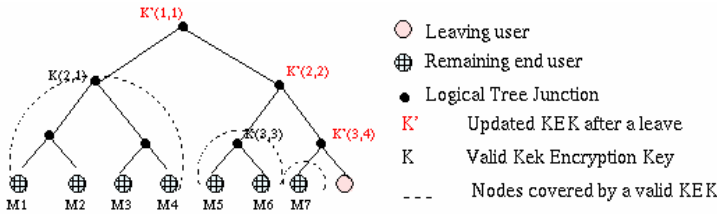


Fig. 2. Example of LKH leaving

4 Proposed Multicast Fingerprinting Scheme

Next we propose a multimedia multicast distribution scheme allowing both copyright protection and encryption. The scheme is fully compatible with existing state-of-the-art multicast encryption techniques and, contrary to the proposals discussed in previous sections, it does not rely on pre-processing techniques nor trusted intermediate nodes.

In our proposal, all multimedia content can be on-line encrypted using symmetric key techniques and updated and managed using simple Logical Key Tree algorithms. All the functionalities for fingerprinting codeword embedding (generating different copies for different end-users), are moved to the end-user side.

4.1 Deployment and Operation

Next we discuss the operational features of the proposed scheme. The overall process consists in the following 8 steps. See also Figure 3:

1. End-user i requests some type of content from the Distributor and joins a multicast group.
2. The Distributor sends all the keys (KEK's and Ks) that the trusted soft-engine i will need to decrypt the content¹
3. The Distributor notifies the trusted soft-engine i about the TTP that will deliver the fingerprinting codeword.
4. The recipient contacts the TTP asking for the fingerprinting codeword and notifies the identity of the Distributor.
5. The TTP delivers the fingerprinting mark to the trusted soft-engine i , and a proof of it (a hash function of the whole mark) to the Distributor. These messages are interchanged using a secure channel.
6. At the accorded time, the distributor starts sending the encrypted content to the multicast group.
7. Upon receiving the encrypted content the trusted soft-engine decrypts and marks it.
8. The marked content is the one that appears on the player.

The system behaviour can be more deeply understood by following the next example. Suppose we want to transmit multimedia content through a multicast channel. As mentioned before, our system needs the interaction of three communication entities to achieve delivery plus secrecy and copyright protection of the multicasted content: the content distributor, the content recipients and a TTP.

When an internet user wants to subscribe the near video on demand service of the distributor (say, become a content recipient) he first accesses the distributor storefront (usually a webpage offering different multimedia products) and asks for the content. As the multimedia product will be delivered using multicast, once the service is required the content recipient needs to join a multicast group.

The Distributor aims to deliver the content in a convenient manner, so its privacy, integrity and authenticity are preserved. This goal can be accomplished by means of well known cryptographic techniques (i.e., encrypting the contents using the AES algorithm). Obviously, the session key must be known by both the distributor and all

the trusted soft-engines that are receiving the content. In this second step, the distributor gives to the recipient the whole set of keys that he will need to decrypt the data.

If only one key (session key) would be given, there would be no way to efficiently update this key. When the session keys shall be updated (every time a recipient joins or leaves the multicast group), the distributor will send the mandatory updating key messages, that once received by the trusted soft-engines

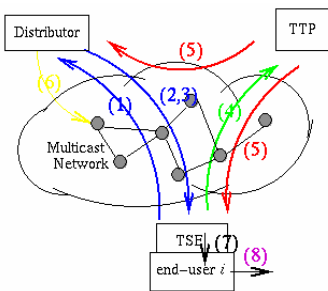


Fig. 3. Atomic steps in the proposed scheme

¹ In case that some key updating needs to be performed, the responsibility of all key management operations completely lies on the Distributor.

will allow them to obtain the new session key. This process is usually done using the LKH algorithm. That is the reason why LKH KEKs are also delivered to the receiver at step 2.

Apart from decrypting the content, the trusted soft engine has the responsibility of embedding the fingerprinting codeword to the displayed stream. To protect the recipient from dishonest distributors, this codeword has to be provided by a TTP trusted by both the distributor and the recipient. If the distributor itself decided the codeword to be inserted in the stream, he could dishonestly accuse of piracy an honest recipient by simply distributing different copies with the same codeword. That is the reason why, in step 3, the distributor tells the recipient where to get the certified mark.

In step 4, the recipient asks the TTP for a valid codeword that uniquely identifies himself. As the TTP has also to notify information about the codeword to the distributor, the recipient also gives to the TTP the identity of the distributor he is dealing with.

After that, the TTP embeds the fingerprinting codeword to the i -th end user trusted soft-engine in order to allow it to properly mark the content. In this step, the TTP also hashes the mark and sends it to the distributor. This value will be used by the distributor as the end-user i node KEK. By means of that, the distributor can be sure that only if the mark is correctly inserted in the soft-engine, the bitstream can be decrypted. If the mark is not inserted in the soft engine, the user i KEK is not known by the soft engine and as a result, the session key won't be able to be recovered.

All these actions have to be performed before the delivery of the multimedia content. When all these steps are finished, the distribution of the content can start. Traditionally, in near-on-demand video services, the content distribution is scheduled at predefined times and users can subscribe for any of those time-slots.

At this point, only the decryption and branding of the bitstream are left to be done. This is entirely the responsibility of the soft-engine and only after this step (7) had been accomplished, the player in the end-user machine can display the decrypted and marked content.

4.2 Marking Process

Now we tackle the problem of embedding the fingerprinting codeword that lies inside the trusted soft-engine into the content that the engine receives.

We assume that the delivered content includes synchronizing information. This means that the player (or decoder) knows at what time has to display each part of the content with respect to the start time. To make the description more clear we will use an MPEG movie delivery as an example. Note that the set of positions that contains the mark is the same for every delivered movie.

We show now how our scheme satisfies the above constrain in a very natural way. When the multicast delivered MPEG sequence first enters the trusted soft-engine, a timer is started. This timer expires after a previously defined time T_g , that we call guard time. At T_g the trusted soft-engine starts buffering the MPEG bitstream, until a marking threshold M_{th} is reached. This threshold is determined by the watermarking algorithm.

The value of the threshold is determined by both the length of the fingerprinting codeword (which in turn depends on the number of end-users) and the watermarking algorithm. Note that in general, different watermarking algorithms will require

different thresholds. Also note that for a given watermarking algorithm and a given threshold we will have a different buffer size, depending on the received bitstream.

At this point, the watermarking algorithm is applied to the buffered MPEG bitstream, and the fingerprinting mark is embedded. The obtained output bitstream uniquely identifies the trusted soft-engine (and obviously the subscribed end-user). This procedure is shown in Figure 4.

Note that the same mark will appear in several parts of the content, thus allowing for a given user to join the group at any time during transmission.

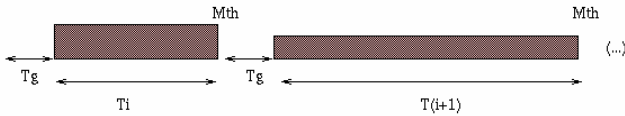


Fig. 4. Guard times and buffering times in the marking algorithm

5 Conclusions

Multicast fingerprinting schemes have been seldom studied due to the difficulty of generating different marked copies in a multicast delivery scenario. Moreover, considered together with encryption, state-of-the-art multicast fingerprinting techniques are difficult to implement.

In this paper we have presented a practical, ready-to-implement solution to the multicast fingerprinting problem. It is based on the existence of what we have called the “trusted soft-engine” a hardware/software device embedded in the end-user side and co-managed by the distributor, the content recipient and a trusted third party. This engine has the responsibilities of handling key updating, obtaining and decrypting the content using the session key, and embedding into the content the mark that uniquely identifies the soft-engine.

The presented scheme is fully compatible with almost all existing key management algorithms for multicast and accepts several fingerprinting codes, depending on the required security level.

Acknowledgements

This work has been partially supported by the Spanish Research Council under SECONNET project (CICYT - TSI2005-07293-C02-01).

References

1. D. Balenson, D. McGrew, A. Sherman. Key Management for Large Dynamic Groups: One-Way Function Trees and Amortized Initialization. *IETF Internet Draft. irtf-smug-groupkeymgmt-oft-00.txt*, work in progress, Aug 2000.
2. A. Barg, G. R. Blakley, G. Kabatiansky, Digital fingerprinting codes: Problem statements, constructions, identification of traitors, *IEEE Trans. Inform. Theory*, 49, 4, 852--865, 2003

3. D. Boneh, J. Shaw, Collusion-Secure Fingerprinting for Digital Data, *Advances in Cryptology-Crypto'95, LNCS*, 963, 452--465, 1995
4. D. Boneh, J. Shaw, Collusion-Secure Fingerprinting for Digital Data, *IEEE Trans. Inform. Theory*, 44, 5, 1897--1905, 1998
5. S. I. Brown, C. Perkins, J. Crowcroft. Watercasting: Distributed Watermarking of Multicast Media. *First International Workshop on Networked Group Communication (NGC '99)*, Pisa, November 17-20, 1999.
6. R. Canetti, T. Malkin, K. Nissim. Efficient Communication Storage Tradeoffs for Multicast Encryption. *Eurocrypt'99* p 456-470 1999
7. R. Canetti, B. Pinkas, A taxonomy of multicast security issues, *IETF Internet Draft draft-irtf-smugtaxonomy-01.txt*, work in progress, April, 1999.
8. G. Caronni, C. Schuba. Enabling Hierarchical and Bulk-Distribution for Watermarked Content. *The 17th Annual Computer Security Applications Conference*, New Orleans, LA, December 10-14, 2001.
9. Chor et al. Tracing Traitors, *IEEE Transactions on Information Theory*, Vol 46, No 3, pp 893-910. 2000
10. H. Chu , L. Qiao, K. Nahrstedt. A Secure Multicast Protocol with Copyright Protection. *Proceedings of IS&T/SPIE Symposium on Electronic Imaging: Science and Technology*, January 25-27, 1999.
11. M. Fernandez, M. Soriano. Soft-Decision Tracing in Fingerprinted Multimedia Contents. *IEEE Multimedia*, vol 11, no2, 2004.
12. T. Hardjono, B. Cain, N. Doraswamy, A Framework for Group Key Management for Multicast Security, IETF Internet Draft, draft-ietf-ipsec-gkmframework-03.txt, work in progress, August 2000
13. H. Harney, E. Harder. Logical Key Hierarchy Protocol (LKH). IETF Internet Draft, harney-sparta-lkhp-sec-00.txt, work in progress, Mar 99.
14. H. Harney , C. Muckenhirn, T. Rivers, Group Key Management (GKMP) Architecture, SPARTA, Inc., IETF RFC 2094, July 1997.
15. P. Judge, M. Ammar. WHIM: Watermarking Multicast Video with a Hierarchy of Intermediaries. *The 10th International Workshop on Network and Operation System Support for Digital Audio and Video*, Chapel Hill, NC, June 26-28, 2000.
16. J. Pegueroles, W. Bin, M. Soriano, F. Rico-Novella. Group Rekeying Algorithm using Pseudo-Random Functions and Modular Reduction. *Grid and Cooperative Computing (GCC)*. Springer-Verlag Lecture Notes in Computer Science, vol 3032, 2004. pp. 875-882
17. N. Wagner. Fingerprinting. *Proceedings of the 1983 IEEE Symposium on Security and Privacy*, pp 18--22, April 1983.
18. D. Wallner, E. Harder, R. Agee, Key management for multicast: Issues and architectures, IETF RFC 2627, June 1999.
19. Yoshioka et al., Systematic Treatment of Collusion secure codes: security definitions and their relations, LNCS 2851, pp 408-421.

Audit-Based Access Control in Nomadic Wireless Environments

Francesco Palmieri and Ugo Fiore

Federico II University, Centro Servizi Didattico Scientifico, Via Cinthia 45,
80126 Napoli, Italy
{fpalmieri, ufiore}@unina.it

Abstract. Wireless networks have been rapidly growing in popularity, both in consumer and commercial arenas, but their increasing pervasiveness and widespread coverage raises serious security concerns. Client devices can potentially migrate, usually passing through very light access control policies, between numerous diverse wireless environments, bringing with them software vulnerabilities and possibly malicious code. To cope with this new security threat we propose a new active third party authentication, authorization and audit/examination strategy in which, once a device enters an environment, it is subjected to security analysis by the infrastructure, and if it is found to be dangerously insecure, it is immediately taken out from the network and denied further access until its vulnerabilities have not been fixed. Encouraging results have been achieved utilizing a proof-of-concept model based on current technology and standard open source networking tools.

Keywords: Mobile networking, security audit, access control.

1 Introduction

The world is going mobile. Today, most consumers take for granted the ability to communicate with friends, family and office anywhere, anytime, at a reasonable cost, so the widespread adoption of low-cost wireless technologies such as WiFi and Bluetooth wireless coverage makes nomadic networking become more and more common. But, as millions of mobile users migrate between home, office, hotel and coffee shop, moving from one wireless access point to the next, they take with them not only their computer, but also electronic hitchhikers they may have picked up in the local shopping mall and unpatched or misconfigured applications. Continual migration from one access point to another with these vulnerabilities threatens the integrity of the other environments, as well as that of other peers within the environments. As corrupted machines move from network to network, they will be able to quickly spread offending code to network resources and users. The traditional paradigm, based on the separation between a secure area inside and a hostile environment outside can no longer be effective, because authenticated users can bring uncontrolled machines in the network. A user may unwittingly bring in active threats such as viruses, Trojan Horses, denial-of-service daemons, or even create a hole for a human intruder; alternatively they may bring in passive threats such as vulnerable packages or poorly

configured software. Particularly, resourceful worms could use nomadic trends to attack and quickly spread in dense urban centers or large campuses, without resorting to the Internet. We believe that this is a fundamental security threat that must be addressed. To contain or at least mitigate the impact and spread of these attack vectors, wireless access control environments must be able to examine and evaluate clients on their first access to the network for potential threats or vulnerabilities. Accordingly, we propose a new third party authentication, authorization and audit/examination paradigm in which once a device enters an environment it is subjected to active analysis by the infrastructure, and if its security level is found to be not adequate it is immediately taken out from the network and denied for any further access until its problems and vulnerabilities have not been fixed. The essence of the proposed architecture is allow or deny access to the network to systems based on their perceived threat, i.e., before they can become infected or attack others. Hence, to evaluate their health status, an active vulnerability assessment is carried out, followed by a continuous passive monitoring against further suspect activities such as port scans, broadcasts or other hostile activities. Our approach seems to be promising for several reasons. Firstly, it does not rely only on global network traffic monitoring. Secondly, it does not require the use of specific agents to be installed on the mobile hosts, implying an undesired strong administrative control on the networked systems, so it is very suitable for the untrusted mobile networking environment. Thirdly, the whole security arrangement is easily applicable at any network location and is not dependent from specific software packages. It would also be effective against unknown viruses, worms or generic malware. Finally, this type of infrastructure would strongly encourage active and timely patching of vulnerable and exploited systems, increasing overall network security. It would benefit users by protecting their systems, as well as keeping them up to date, and benefit local providers by protecting their infrastructure and reducing theft of service. Deployment would also protect the Internet as a whole by slowing the spread of worms, viruses, and dramatically reducing the available population of denial-of-service agents.

The remainder of this paper is organized as follows. Section 2 reviews the related scientific literature. Section 3 briefly sketches the ideas behind the proposed architecture. The detailed components of the overall system and their interaction are described in Section 4. Section 5 presents a proof-of-concept implementation realized with current technology. Finally, Section 6 is dedicated conclusions and planning for future research.

2 Related Work

IEEE standard 802.11i [1] is designed to provide enhanced security in the Medium Access Control (MAC) layer for 802.11 networks. While the standard provides good coverage of confidentiality, integrity, and mutual authentication when CCMP is used, active research is devoted to identifying potential weaknesses and suggesting security improvements. Many researchers are concentrated on discovering weaknesses in association and authentication, and in particular in the 4-Way Handshake procedure (see, for example [2]). Much effort has also been spent on DoS attacks. Management frames are unprotected, thus even an adversary with moderate equipments can easily forge these frames to launch a DoS attack. Ding *et al.* [3] proposed the use of a

Central Manager (CM) to dynamically manage Access Points and their clients. Another approach is to respond to deauthentication and disassociation frames by triggering a new 4-Way Handshake [4], while Ge and Sampalli [5] suggest that the management frames should be authenticated. He and Mitchell [6] described some attacks and proposed corresponding defenses. They also compiled a survey of the relevant literature. All these works are targeted at hardening the protocols and denying network access to unauthorized devices, while there appears to be much less work focused on guaranteeing network protection by proactively identifying potential threats even if they come from legitimate clients. Our scheme addresses the problem of authenticated users bringing vulnerable or infected machines into a secure network.

3 The Basic Architecture

Information security officers and systems and network administrators want to control which devices are allowed onto their networks. In academic environments this is even more important, because university networks are usually well connected to the Internet and are therefore popular targets for hackers. In addition many systems in these environments are owned by people nearly completely unaware of security chores as keeping their operating systems or anti-virus definitions up-to-date. The fundamental concept of the proposed active access control strategy is that once a client mobile device enters a new environment it has to be subjected to analysis by the infrastructure to examine for potential vulnerabilities and contaminants and evaluate its security degree to decide if network access can be granted without compromising the target environment. This enhanced security facility can be implemented as an additional access control phase immediately following the traditional 802.11i authentication and authorization paradigm. Here, the involved wireless Access Point after performing successful 802.1x authentication of the supplicant against the RADIUS or DIAMETER server and when the mobile node receives full network access and is reachable through an IP address, requires its complete security analysis by a new network entity that we will call the *Auditor*. To ensure the proper model scalability in huge and complex network environments there may be many Auditors associated to the different access points, each one, and eventually more than one, dedicated to the coverage of a specific area. This also provides fault tolerance and implicit load distribution where necessary.

There are a large set of possible types of analyses that could be performed by the Auditor, including external network scans and probes of offered services, virus scans, or behavior monitoring. For instance, a simple type of examination might determine the versions of installed OS and software and appropriate security patches, verifying the existence of open vulnerabilities where applicable. During the different examination phases, the Auditor computes a complex score representing the resulting client insecurity degree, built upon the sum of individual scores assigned to the different analysis results. The task of specifying the partial scores to be assigned to different categories is left to the network administrator, because environments and security requirements may vary widely. Once the security analysis terminates, the above degree is compared with an acceptance threshold configured at the auditor level and, if the computed security score does not fall below the threshold, the auditor returns a reject response to the access points that immediately denies any further network access to the examined device. Otherwise, the mobile client don't loose its network access, but

the examination will not stop for the involved node but passes from the active to the passive state, which means that using standard intrusion detection techniques, the auditor, on behalf of the local infrastructure, could continuously examine network traffic to determine if any entity is trying to launch a port scan or any other form of attack or take over other machines. In this case the Auditor quickly notifies the above event to the access point and consequently the node is immediately taken off from the network. The whole process is synthesized in the status diagram below.

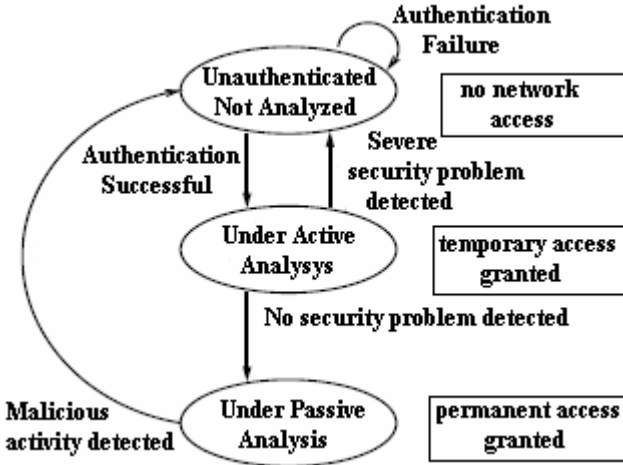


Fig. 1. The audit-based access control status diagram

The proposed architecture can scale quite easily. It can be adapted to a large network by simply replicating its components. All that is needed is to properly set up relationships between each Auditor and its controlled access points. The resulting system is flexible, customizable, extensible, and can easily integrate other common off-the-shelf tools. In addition, it provides automated support for the execution of complex tasks that require the results obtained from several different tools.

4 The Operating Scenario

Available technologies implement wireless security and authentication mechanisms such as 802.1X and 802.11i to prevent the possibility of unauthorized users gaining undue access to a protected network. Security is typically provided via access control. Unfortunately, this arrangement falls short as a tool to avoid the inside spread of hostile agents, since in the modern nomadic networking environment authentication and authorization are no longer sufficient and need to be completed with a strong security assessment aiming at clearly highlighting any potential menace. Therefore our proposed security strategy operates according to the scenario described below.

Four entities are involved, called the Supplicant (the wireless station), the Authenticator (the Access Point), the Authentication Server, and the Auditor. We assume, as in IEEE 802.11i, that also the AP and Auditor have a trustworthy UDP channel

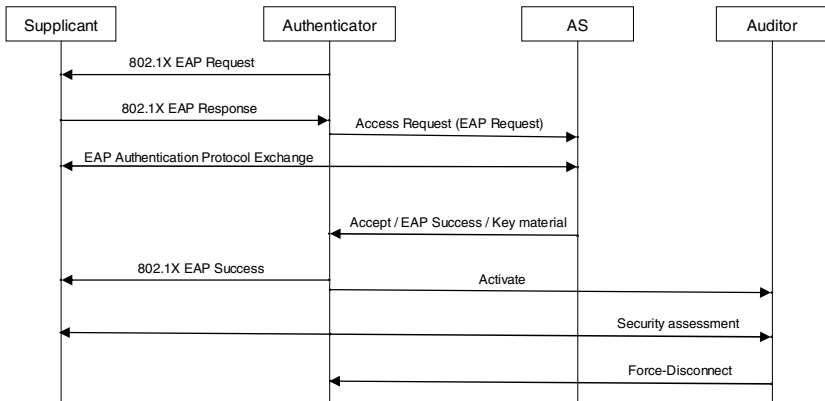


Fig. 2. Operating scenario

between them that can be used to exchange information and action triggering messages, ensuring authenticity, integrity and non-repudiation.

After successful authentication, when a wireless device is granted access to the network, it usually issues a DHCP request for an address. The local DHCP server then hands an IP address to the wireless device passing through the access point that has been modified to detect the DHCP response message (message type DHCPACK) and acquire knowledge about the layer-3 availability of a new device together with its IP address. Now the access point has to ask the Auditor for the needed security analysis.

4.1 Message Layout

Basically, the communication between the Auditor and the Access Point is twofold: the Access Point signals the Auditor that a new Client has associated and that the assessment should begin; the Auditor may request that the Access Point disconnect the Client in case the computed score exceed the admissibility threshold. So only two messages are necessary, respectively the Activation message and the Force-Disconnect message, whose layout is shown in the following figure.

	Octet Number
Protocol Version	1
Packet Type	2
Packet Body Length	3-4
Identifier	5-8
Client MAC Address	9-14
Client IP Address	15-18

Fig. 3. Message layout

Port UDP/7171 is used on both the Auditor and the Access Point to exchange messages with the other party. We also preferred to leave room for further extensions. For instance, message authentication may be called upon, as there is potential for DoS attacks if such messages can be forged.

- **Protocol Version.** This field is one octet in length, taken to represent an unsigned binary number. Its value identifies the version of protocol supported by the message sender.
- **Packet Type.** This field is one octet in length, taken to represent an unsigned binary number. Its value determines the type of packet being transmitted. The following types are defined:
 - **Activate.** A value of 0000 0000 indicates that the message carries an Activate packet.
 - **Force-Disconnect.** A value of 0000 0001 indicates that the message is a Force-Disconnect packet.
- **Packet Body Length.** This field is two octets in length, taken to represent an unsigned binary number. The value of this field defines the length in octets of the packet body following this field; a value of 0 indicates that there is no packet body.
- **Identifier.** The Identifier field is one octet in length and allows matching of responses with requests.
- **Client MAC Address.** This field is six octets in length and specifies the MAC address of the Client.
- **Client IP Address.** This field is four octets in length and indicates the IP address of the Client.

No acknowledgement messages have been provided for, as the communication between the Access Point and the Auditor is asynchronous and neither actor needs information about the completion of task by the other. Synchronous communication would induce unacceptable latency.

4.2 The Security Analysis Process

The security analysis process should take advantage of a reliable and possibly complete knowledge base that contains semantically-rich information provided by network mapping and configuration assessment facilities. Accordingly, client profiling will be accomplished through the use of specialized network analysis tools that can be used to examine open ports and available services in a fairly non-intrusive manner. This analysis can detect the presence of services that are unnecessary or undesirable in the given environment or identify explicit anomalies and system vulnerabilities. For example, if a port scan was run and determined that a normally unused port, e.g., TCP port 1337, was open on a scanned host, a penalty would be added to the security score indicating that the machine had been potentially exploited and may represent a security problem. Our philosophy is to use only open source applications, in order to avoid licensing costs and legalities, and perform each part of the overall service discovery/analysis task with the tool best suited for the job. This works best if a tool performs one specific task instead of implementing many different functionalities as a monolithic application. A tool that is able to perform many tasks should at least have parameters that can limit the operation of the tool to exactly what is needed.

Examples of these tools are Nessus [7] and Nmap [8] both falling under the GPL (General Public License) and configurable by fine tuning the proper options to suit the specific needs of each analysis' task. The advantage of using such existing tools is that it requires less work to implement the client examination procedures.

The first phase of the examination, the information gathering phase, comprises the determination of the characteristics of the target mobile node such as OS type, and "listening" applications e.g. WWW servers, FTP services, etc. This is ordinarily achieved by applying the following techniques:

- **OS detection:** A common OS detection technique is "IP stack fingerprinting" – the determination of remote OS type by comparison of variations in OS IP stack implementations behavior. Ambiguities in the RFC definitions of core internet protocols coupled with the complexity involved in implementing a functional IP stack enable multiple OS types (and often revisions between OS releases) to be identified remotely by generating specifically constructed packets that will invoke differentiable but repeatable behavior between OS types, e.g. to distinguish between Sun Solaris, Microsoft Windows NT and the most common Open Source Unix flavors such as Linux and *BSD. By determining the presence of older and possibly unpatched OS releases the Auditor can increase the overall insecurity score according to the locally configured policy. For example, if network administrators know that all of the legitimate clients are running Windows boxes, they will likely raise the score of an OS fingerprint indicating Linux, or the reverse. In those cases when OS fingerprinting is able to determine the operating system version, besides the type, higher scores can be assigned to the older versions. Multiple tools, such as nmap, queso and p0f have been used together, for results reliability sake, to implement OS fingerprinting in our architecture.
- **Service Detection:** a.k.a. "port scanning" performs the availability detection of a TCP, UDP, or RPC service, e.g. HTTP, DNS, NIS, etc. Listening ports often imply associated services, e.g. a listening port 80/tcp often implies an active web server. The presence of deprecated services, from the security point of view, or, at worst port associated to possible worm infection. Here, for each dangerous service or suspicious port detected the insecurity score is increased according to a configurable policy, such as the example one sketched in the table below. Furthermore, the pattern of listening ports discovered using service detection techniques may also indicate a specific OS type; this method is particularly applicable to "out of the box" OS installations. Service detection will be accomplished with the nmap tool that can provide a very fast and complete assessment of the running services and in some cases their versions [8].
- **Vulnerability detection:** Based on the results of the previous probe, that attempts to determine what services are running, a new and more sophisticated scanner, with semantic knowledge about up-to-date security issues, tries to exploit known vulnerabilities of each service in an attempt to test the overall system security. From the implementation point of view, the results of the previous nmap scan can be used by a more sophisticated tool, such as Nessus [7] to conduct a more directed and thorough assessment. Nessus, offers much more advanced configuration audit functionalities, has the capability of remotely scanning a client to determine running

services, the versions, and if the client is susceptible to specific security threats. The assessment library associated with Nessus is very comprehensive, covering a large variety of architectures, operating systems, and services. After the assessment, each detected problem heavily taxes the computed insecurity score since it is typically a very dangerous symptom or an open vulnerability, whereas the presence of open ports outside an admissible range specified by the network administrator could be evaluated differently, in accordance with locally defined policies.

Assessment should be logged and their result retained for a time interval configurable by the network administrator, in order to prevent a Client continuously trying to get access from consuming inordinate amounts of resources. The optimal value of this amount of time is a tradeoff between the quest for accurate and up-to-date security assessment and the compelling need to save resources.

The Auditor also acts as a passive unfriendly activity detector. This is a necessary facility, since the analysis can take more time than can be acceptable for mobile users, and so we decided to grant immediate access to users, without waiting for the assessment completion. This opens a security risk, because, a mobile client can immediately be active to propagate malware, but this activity would be immediately noticed by the passive detector. In fact, if, at any time, some hostile activity such as scanning, flooding or known attacks is discovered from an associated Client, the Auditor requests the AP to disconnect it. Each Auditor is assigned two IP addresses. The first is utilized to communicate with the connected Access Points. The other will be dedicated to passive listening to signal illegitimate activity. To detect the above misbehavior, a simple but very flexible network intrusion detection system (NIDS) such as **snort** (open source intrusion detection system) [9] can be used. It performs real-time traffic and protocol analysis, signature matching against a signature database (kept constantly auto-update via Oinkmaster services) and packet logging.

5 Implementation

A simple proof-of-concept system was developed to test the effectiveness of the proposed system, with an emphasis on the use of currently available devices and open-source components. The Authenticator has been implemented on a modified Linksys Access Point, starting from its publicly available Linux-based firmware. Here we introduced the following new functions:

- Accept, decode and verify the new protocol messages
- Detect the DHCP ACK message and send the Activation message
- Disconnect the client upon reception of a Force-Disconnect message

On the other side, the Auditor has been implemented on a simple Intel 486-based Soekris [10] appliance running OpenBSD [11] with **nmap**, **nessus**, **queso**, **p0f** and **snort** applications installed. Although this implementation it is only a “*proof of concept*”, our simple testbed demonstrated the correct operation of the proposed security framework.

6 Conclusions and Future Work

In this paper, we address the problem of authenticated users bringing vulnerable or infected machines into a secure network. In a nomadic wireless environment where machines are not under the control of the network administrator, there is no guarantee that operating system patches are correctly applied and anti-malware definitions are updated. We propose an architecture based on a security assessment to detect some characteristics of the inspected host and identify vulnerabilities. The proposed architecture can scale quite easily. It can be adapted to a large network by simply replicating its components. All that is needed is to properly set up relationships between each Auditor and its controlled Access Points.

Areas for future research and development include better user interaction. Disconnection, and its reason, should be explicitly notified to users so that they would know what happened and the appropriate actions required to fix the problem. If a downright disconnection is felt to be too draconian and some restricted environment should be created, a tighter integration with DHCP would also be called for.

References

1. IEEE 802.11i, *Medium Access Control (MAC) Security Enhancements*, Amendment 6 to IEEE Standard for Information technology – Telecommunications and information exchange between systems – Local and metropolitan area networks – Specific requirements – Part 11: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications, July 2004.
2. Mishra, A., and Arbaugh, W. A., *An initial security analysis of the IEEE 802.1X standard*. Technical Report CS-TR-4328, UMIACS-TR-2002-10, University of Maryland, February 2002
3. Ding, P., Holliday, J., and Celik, A., *Improving the security of Wireless LANs by managing 802.1X Disassociation*. In Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC '04), Las Vegas, NV, January, 2004.
4. Moore, T. *Validating 802.11 Disassociation and Deauthentication messages*. Submission to IEEE P802.11 TG1, September, 2002
5. Ge, W., Sampalli, S., *A Novel Scheme For Prevention of Management Frame Attacks on Wireless LANs*, March 2005, <http://www.cs.dal.ca/news/def-1341.shtml>
6. He, C., and Mitchell, J., *Security Analysis and Improvements for IEEE 802.11i*, in 11th Annual Network and Distributed System Security Symposium (NDSS '05). San Diego, February 2005.
7. Nessus Security Scanner (<http://www.nessus.org/>).
8. Nmap Security Scanner (<http://www.insecure.org/>).
9. Snort, Open Source Network Intrusion Detection System (<http://www.snort.org/>).
10. Soekris Engineering (<http://www.soekris.com/>).
11. Open BSD (<http://www.openbsd.org/>).

Cost – Time Trade Off Models Application to Crashing Flow Shop Scheduling Problems

Morteza Bagherpour¹, Siamak Noori², and S. Jafar Sadjadi²

¹ Ph.D. student of Industrial Engineering department,
Iran University of Science & Technology, Narmak, Tehran, Iran
mortezaabagherpour@gmail.com

² Department of Industrial Engineering,
Iran University of Science & Technology, Narmak, Tehran, Iran
snoori@iust.ac.ir,
sjsadjadi@iust.ac.ir

Abstract. Many traditional cost– time trades off models are computationally expensive to use due to the complexity of algorithms especially for large scale problems. We present a new approach to adapt linear programming to solve cost time trade off problems. The proposed approach uses two different modeling flowshop scheduling into a leveled project management network.

The first model minimizes makespan subject to budget limitation and the second model minimizes total cost to determine optimum makespan over production planning horizon.

Keywords: Flow shop scheduling, linear programming, Leveled project management network, Makespan.

1 Introduction

In an m - machine flow shop problem, there are n -jobs to be scheduled on m -machines where each job consists of m -operations and each operation requires a different machine and all jobs are processed in the same order of the machines.

Extensive research has been conducted for variants of the regular flow shop problem with the assumption that there exists an infinite buffer space between the machines.

Even though such an assumption is valid for some applications, there are numerous situations in which discontinuous processing is not allowed. Such flow shops are known as no-wait. A no-wait flow shop problem occurs when the operations of a job have to be processed continuously from start to end without interruptions either on or between machines. This means, when necessary, the start of a job on a given machine is delayed in order that the operation's completion coincides with the start of the next operation on the subsequent machine. There are several industries where the no-wait flow shop problem applies. Examples include the metal, plastic, chemical and food industries. For instance, in the case of rolling of steel, the heated metal must continuously go through a sequence of operations before it is cooled in order to prevent defects in the composition of the material. Also in the food processing

industry, the canning operation must immediately follow the cooking operation to ensure freshness. Additional applications can be found in advanced manufacturing environments, such as just in time and flexible manufacturing systems.

Sequencing methods in the literature can be broadly categorized into two types of approaches, namely optimization and heuristic. Optimization approaches guarantee to obtain the optimum sequence, whereas heuristic approaches mostly obtain near-optimal sequences. Among the optimization approaches, the algorithm developed by Johnson (1994)[6] is the widely cited research dealing with sequencing n jobs on two machines. Lomnicki (1965)[7] proposed a branch and bound technique to find the optimum permutation of jobs. Since the flow shop scheduling problem has been recognized to be NP-hard, the branch and bound method cannot be applied for large size problems. This limitation has encouraged researchers to develop efficient heuristics. Heuristics are two-fold: constructive and improvement. In constructive category, methods developed by Palmer (1965) [3], Campbell et al.(1970) [2], Gupta (1971)[4], Dannenbring (1977)[12], RKock and Schmidt (1982) [13] and Nawaz et al. (1983) [8] can be listed. Mostly, these methods are developed on the basis of the Johnson's algorithm. Turner and Booth (1987) [15] and Taillard (1990)[10] have verified that the method proposed by Nawaz et al.(1983) [9], namely NEH, performs superior among the constructive methods tested. On the other hand, Osman and Potts (1989)[11], Widmer and Hertz (1989) [16], Ho and Chang (1990)[5], Ogbu and Smith (1990) [10], Taillard (1990) [14], Nowicki and Smutnicki (1996)[8] and Ben-Daya and Al-Fawzan (1988) [16] have developed improvement heuristics for the problem of interest. Also Aldowaisan and Allahverdi (2004) developed new heuristics for m -machine no-wait flow shop to minimize total completion time.

2 Flow Shop Scheduling

There are some assumptions in flow shop scheduling problems as follows:

2.1 Assumption Regarding the Jobs

- The sequencing model consists of a set of n jobs, which are simultaneously available.
- Job j has an associated processing time, which is known in advance on machine i .
- All machines are continuously available.
- Set-up time is independent of sequence and is included in the processing time.
- Jobs are independent of each other.
- Each job, once started, continues in process until it is completed.

2.2 Assumption Regarding the Machines

- The shop consists of m machines.
- No machine processes more than one job at a time.
- Each machine is available at all times for processing jobs without any interruption.
- Each machine in the shop operates independently.
- Machines can be kept idle.

2.3 Assumption Regarding the Process

- Processing time for each job on each machine is deterministic and independent of the order in which jobs are to be processed.
- Transportation times of a job between machines and set-up times are included in the order of job processing.

3 Converting Flow Shop Scheduling into Leveled Project Management Network

In flow shop problem there are m machines and n jobs, which is assumed that sequence of the jobs on the machines is to be known. As we are interested in reducing make span by employing additional resources, method developed for crashing the project network can be used if flow shop can be represented as a project network. To avoid any resource overlapping it is needed to level the resources.

We propose a new approach to reduce makespan using additional resources. The proposed approach converts a flow shop problem into a leveled project management network.

3.1 Network Requirements

There are some network requirements needed to hold when a flow shop problem is converted into a network such as nodes, activities, etc as follows:

Nodes: Each node represents an event of starting or finishing operation(s) on machines.

Activity: Activities are the operations to be proceeding on specific machine and have duration equal to processing time.

Predecessors: For any activity representing the previous operation for the same job constitutes a preceding activity to the operation. Further the activity corresponding to the operation of the job on the same machine, which is before this job in the sequence also constitutes preceding activity.

Table 1. Data of flow shop problem

Activity	Predecessors	Duration time	Machine
1,1	-	D_{11}	M_1
.	.	.	.
.	.	.	.
i,j	$(i-1,j),(i,j-1)$	D_{ij}	M_i
.	.	.	.
.	.	.	.
m,n	$(m-1,n),(m,n-1)$	D_{mn}	M_m

Duration time: "processing time" is the duration of the activity.

Resources: Machines are the only resources.

Suppose we have a flow shop system with n jobs and m machines. Table 1 describes the specifications:

3.2 Crashing the Network

Different linear programming models – depending type of objectives is used to crash the network. In next section two type of linear programming formulations is explained where the selection of crashed operations is optimized due to objectives.

4 The Proposed Approach

Notations:

i-j: activity from node i to node j

D_n (i-j): normal duration time of activity from node i to node j

D_f (i-j): minimum crash time of activity i-j

d_{ij}: crashed activity time for activity i-j

C_{ij}: crash cost per unit time for activity i-j

e :number of network events

t_i: planning date for event i

B: Pre-specified budget

H: Overhead cost

K: Fixed cost

4.1 Model 1: Minimizing Makespan Subject to Budget Limitation

In model 1, the objective is to determine the operations to be crashed in order to find the minimum make span subject to budget limitation. When we convert flow shop scheduling problem into a network, it is possible to crash the network to find minimum make span. This problem can be formulated as follows:

$$\text{Min } Z = (t_n - t_1)$$

subject to:

$$t_j - t_i \geq d_{ij} \tag{1}$$

$$D_f(i-j) \leq d_{ij} \leq D_n(i-j) \tag{2}$$

$$\sum \sum C_{ij} (D_n(i-j) - d_{ij}) \leq B \tag{3}$$

$$t_i \geq 0, d_{ij} \geq 0$$

The objective function of model 1 is to minimize additional cost for crashing the activities. Constraint (1) states that the start time of event j should be at least equal to start time of i and crash duration of activity i-j. Constraint (2) is related to the lower and upper bounds on crash duration and finally constraint (3) is for budget limitation. Note that all *t_i*'s and *d_{ij}* must be non-negative and also integer. However, due to the structure of the problem, model 1 can be solved as a linear programming problem.

4.2 Model 2: Minimizing Total Cost to Determine Optimum Makespan

In model 2 we are interested in optimum common due date when total cost is minimized. Therefore, we convert the flow shop scheduling problem into a leveled

project management network by using the crashing of operations on critical path to attain the objective. Thus, the problem can be formulated as follows:

$$\begin{aligned} \text{Min } Z &= H(t_n - t_1) + \sum \sum C_{ij} (D_n(i-j) - d_{ij}) + K \\ \text{subject to:} \\ t_j - t_i &\geq d_{ij} & (1) \\ D_f(i-j) &\leq d_{ij} \leq D_n(i-j) & (2) \\ t_i &\geq 0, d_{ij} &\geq 0 \end{aligned}$$

Note that all constraints described in model 1 and the objective function is the minimization of total cost over the production planning horizon.

4.3 Converting Network into Flow Shop Scheduling Problem

In the previous section we have shown how to convert flow shop problems to critical path method (CPM) network. We can calculate the earliest start, latest start, floats for all the activities using CPM method. Earliest finish time of the project is equal to make span for the flow shop problem. Then interpretation of results gained by linear programming models into flow shop scheduling problems could be done.

5 Numerical Example

In this section, in order to show the implementation of our proposed models, two numerical examples are given.

5.1 The First Numerical Example

Consider a flow shop problem with three machines and five jobs, where the sequence is obtained as E-C-A-D-B [Campbell et al]. The corresponding processing times are given in table 2 and the overhead cost is equal to 20 per unit and the fixed cost is equal to 100. Critical value is equal to 41. Also pre specified budget is considered equal to 26.

At first the problem is converted to a network where the information of prerequisite is given in table 3 and the corresponding network is shown in figure 1. The white arrows considered as dummy activities.

Table 2. The processing times of a flow shop with 3 machines and five jobs

		Machine		
		M ₁	M ₂	M ₃
J O B	A	8	3	5
	B	4	8	3
	C	7	3	8
	D	8	3	8
	E	3	5	8

Table 3. The Project network data for corresponding example

Activity	Predecessors	Duration Time	Crash Time	Slope of Cost
1,1	-	3	2	10
1,2	1,1	7	6	8
1,3	1,2	8	8	-
1,4	1,3	8	8	-
1,5	1,4	4	3	8
2,1	1,1	5	3	6
2,2	(2,1),(1,2)	3	3	-
2,3	(2,2),(1,3)	3	3	-
2,4	(1,4),(2,3)	3	3	-
2,5	(1,5),(2,4)	8	6	8
3,1	2,1	8	8	-
3,2	(2,2),(3,1)	8	8	-
3,3	(2,3),(3,2)	5	4	12
3,4	(3,3),(2,4)	8	8	-
3,5	(3,4),(2,5)	3	3	-

Table 4 summarizes the optimal solution according to budget limitation.

Table 4. Optimal duration for each activity according to budget limitation

Activity Name	Duration
d1 2	2
d2 3	6
d5 6	4
d2 7	5
d14 18	7
d16 17	5
d3 4	8
d4 5	8
d8 9	3
d17 18	8
d10 11	3
d12 13	3
d7 15	8
d15 16	8
d18 19	3

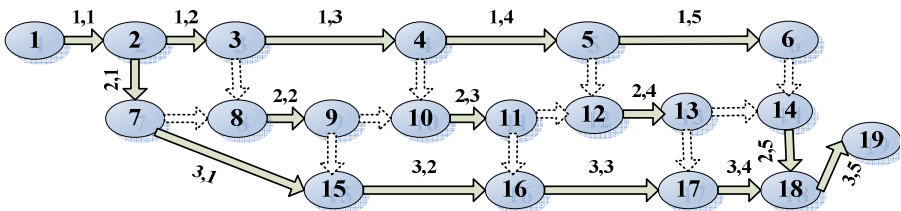


Fig. 1. Project network for given example

5.2 The Second Numerical Example

Consider available data available in pervious example. In table 5, according the objective, optimal duration of each activity is given.

Table 5. Optimal duration of each activity

Activity Name	Duration
d1 2	2
d2 3	6
d5 6	4
d2 7	4
d14 18	7
d16 17	5
d3 4	8
d4 5	8
d8 9	3
d17 18	8
d10 11	3
d12 13	3
d7 15	8
d15 16	8
d18 19	3

As we could observe, with an implementation of a linear programming model, we could easily find an optimal resource allocation for a flowshop problem. Obviously, it is an easy task to apply the proposed method of this paper for relatively large scale projects which are embedded with numerous numbers of activities to be crashed. Although there are many researches which are either on flow shop scheduling, project management network, linear programming applications to crashing network, to the best of our knowledge, there is no closely related research which could help us compare our results.

6 Conclusion and Further Research

In this paper, we have presented two applications for flow shop scheduling problems. The proposed approach converts flow shop problem to a leveled project management network and the optimal makespan has obtained using linear programming models. The objective of the first model is to minimize makespan subject to budget limitation and in the second model the objective is to find common due date to provide better delivery to the customer. However, the proposed approach can be used for large scale projects which are embedded with numerous numbers of activities to be crashed. Further research could be focused on applying linear programming methods to crashing parallel flow shop scheduling where the problem includes k flow shop problems which are working in parallel systems, simultaneously.

References

- [1] Aldowaisan.T, Allahverdi.A, 2004, New heuristics for m-machine no-wait flowshop to minimize total completion time, *International journal of management science*, 32, 345-352.
- [2] Campbell HG, Dudek RA, Smith ML. 1970, A heuristic algorithm for the n-job, m-machine sequencing problem. *Management Science*; 16(B):630–7.
- [3] Dannenbring.DG. 1977, An evaluation of flow shop sequencing heuristics. *Management Science*,23(11):1174–82.
- [4] Gupta JND. 1971,A functional heuristic algorithm for the flow shop scheduling problem. *Operational Research Quarterly*,;22(1):39–47.
- [5] Ho JC, ChangYL. 1990,A new heuristic for the n-job, m-machine flow shop problem. *European Journal of Operational Research*; 52:194–206.
- [6] Johnson SM. 1954, Optimal two- and three-stage production schedules with setup times included. *Naval Research Logistics Quarterly*;1(1):61–8.
- [7] Lomnicki ZA. 1965, A branch and bound algorithm for the exact solution of the three-machine scheduling problem, *Operational Research Quarterly*;16(1):89–100.
- [8] Nowicki E, Smutnicki C. 1996,A fast tabu search algorithm for the permutation flow shop problem. *European Journal of Operational Research*; 91:160–75.
- [9] Nawaz M, Enscoore Jr. EE, Ham I. 1983, A heuristic algorithm for the m-machine, n-job flow shop sequencing problem. *OMEGA: International Journal of Management Science*;11(1):91–5.
- [10] Ogbu FA, Smith DK. 1990, The application of the simulated annealing algorithm to the solution of the n/m/Cmax flow shop problem. *Computers and Operations Research*;17:243–53.
- [11] Osman IH, Potts CN. 1989, Simulated annealing for permutation flow shop scheduling. *OMEGA: International Journal of Management Science*; 17:551–7.
- [12] Palmer DS. 1965,Sequencing jobs through a multi-stage process in the minimum total time—a quick method of obtaining a near optimum. *Operational Research Quarterly*; 16(1):101–7.
- [13] RKock H, Schmidt G. 1982, Machine aggregation heuristics in shop scheduling. *Methods of Operations Research*; 45: 303–14.
- [14] Taillard E. 1990, Some efficient heuristic methods for the flow shop sequencing problem. *European Journal of Operational Research*; 47:65–74.
- [15] Turner S, Booth D. 1987, Comparison of heuristics for flow shop sequencing. *OMEGA: International Journal of Management Science*; 15:75–8.
- [16] Widmer M, Hertz A. 1989,A new heuristic for the flow shop sequencing problem. *European Journal of Operational Research*; 41:186–93.

The ASALB Problem with Processing Alternatives Involving Different Tasks: Definition, Formalization and Resolution*

Liliana Capacho^{1,2} and Rafael Pastor²

¹ Universidad de Los Andes, Dpto. de I.O. - EISULA y CESIMO, Mérida, Venezuela

² Technical University of Catalonia, IOC Research Institute, Av. Diagonal 647, Edif. ETSEIB, p.11, 08028 Barcelona, Spain

{liliana.capacho, rafael.pastor}@upc.edu

Abstract. The Alternative Subgraphs Assembly Line Balancing Problem (ASALBP) considers assembly alternatives that determine task processing times and/or precedence relations among the tasks. Capacho and Pastor [3] formalized this problem and developed a mathematical programming model (MILP) in which the assembly alternatives are determined by combining all available processing alternatives of each existing sub-assembly. In this paper an extended definition of the ASALBP is presented in which assembly sub-processes involving different tasks are also considered. Additionally, a mathematical programming model is proposed to formalize and solve the extended version of the ASALBP, which also improves the performance of the former MILP model. Some computational results are included.

1 Introduction

The classical Assembly Line Balancing Problem (ALBP) consists in assigning a set of tasks to a group of workstations in order to optimize certain efficiency measure (e.g. the number of workstations). Each task is characterized by a processing time and a set of precedence relations, which specifies the allowed processing order.

ALBP are classified into two well-known categories [1]: Simple Assembly Line Balancing Problems (SALBP) and General Assembly Line Balancing Problems (GALBP). SALBP involve very simple and restrictive problems which consider, for example, a unique serial line that processes a single model of one product. GALBP are those in which one or more assumptions of the simple case are varied. Amongst such problems the following groups are usually considered: UALBP that involve U-shaped lines that may be used to overcome the inflexibility of serial lines; MALBP which appear when a line produces units of different models of a unique product (see, for example, Milterburg [8]); MOALBP which consider several optimization objectives; and RALBP that consider robotic lines. Other less common GALBP considered in the literature include problems that involve parallel workstations; parallel tasks; stations not equally equipped, which may imply equipment selection; two-sided or buffered lines; workstation capacity constrained; problems involving processing times that are dependent on the sequence, stochastic or fuzzy; and problems considering multi-product lines (e.g. [5], [9], [12]).

* Supported by the Spanish MCyT project DPI2004-03472 and co-financed by FEDER.

Assembly line balancing problems have been widely studied (e.g. Baybars [1], Becker and Scholl [2], Scholl and Becker [10]). Although most published research work on assembly line balancing focuses on the simple case, in recent years a significant amount of research effort has been directed at the general case in order to address more realistic problems.

Numerous procedures have been developed to solve assembly line balancing problems, which are usually grouped into two main categories: exact methods and approximate methods. According to Becker and Scholl [2], most exact methods are based on linear and dynamic programming and branch and bound procedures (e.g. Scholl and Klein [11]). A wide range of approximate methods have been developed to efficiently solve more realistic ALBP. Among these there are heuristics based on priority rules and enumeration procedures (e.g. Lapiere and Ruiz [7]); metaheuristics such as genetic algorithms (e.g. Kim et al. [6]); simulated annealing (e.g. Suresh and Sahu [12]); tabu search (e.g. Pastor et al. [9]); and others that include procedures based on ant colonies, constrained logic programming, fuzzy logic, expert systems and algorithms based on networks theory (e.g. Gen et al. [5]).

A common feature of ALBP is that a unique and predetermined precedence graph is used to represent all relations among the tasks required to assemble a particular product. In real-life problems, however, is possible that some parts of a product admit several alternative assembly variants that cannot be represented in a standard precedence graph.

In this regard, Capacho and Pastor [3] presented and formalized a new GALBP entitled ASALBP (Alternative Subgraphs Assembly Line Balancing Problem), which considers the possibility of processing alternatives in which the processing time and/or precedence relations of some tasks depend on the selected sub-assembly alternative. In this problem, each processing alternative consists of a particular processing order of a subset of tasks.

Usually, in assembly line balancing processes an alternative for each sub-assembly is selected a priori and the line is then balanced. To solve the ASALBP efficiently, two subproblems need to be solved simultaneously: the decision problem, which selects the assembly sequence of those parts that admit alternatives, and the balancing problem, which assigns the tasks to the workstations.

The remainder of this paper is organized as follows. Section 2 introduces the Alternative Subgraphs Assembly Line Balancing Problem. Section 3 presents the extended definition of the ASALBP. Section 4 describes the model proposed to solve the ASALBP regarding the extended definition. Section 5 provides some computational results. Finally, Section 6 presents the conclusions and proposes future research work.

2 ASALBP: The Alternative Subgraphs Assembly Line Balancing Problem

As previously mentioned, the ASALBP is a new general assembly line balancing problem that considers assembly alternatives, in which an alternative has to be

selected for each part that admits assembly variants. Each processing alternative could be represented by a precedence graph as can be seen in Figure 1, which shows two assembly alternatives for a simple example involving six tasks. In alternative 2 the processing order changes for tasks B, C, D and E. Furthermore, the processing time of task B increases 2 time units when it is processed after task E.

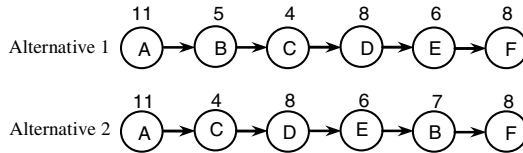


Fig. 1. Two assembly alternatives

Normally, when a problem with assembly alternatives has to be solved one of the available alternatives for each sub-assembly is selected a priori and the line is then balanced. Smallest total processing time is a common criterion that system designers usually use to choose an assembly alternative. If these two processes – the selection process and the line balancing – are carried out independently, it cannot be guaranteed that the global problem can be solved optimally. However, better solutions can be obtained if the two problems are solved simultaneously, as illustrated in Figure 1. If the two resulting balancing problems are solved optimally, minimizing the number of workstations given an upper bound on the cycle time of 15 time units, the results shown in Table 1 are obtained. Even though it has a longer total processing time (44), Alternative 2 provides the best number of workstations for the problem. If the selection process had been carried out a priori, Alternative 1 would have been selected, since its total processing time is 42, and the best solution would have been discarded.

Table 1. Results of balancing the assembly alternatives

Alt	Tasks per station (time)				Total processing time	No. of stations
	I	II	III	IV		
1	A (11)	B,C (9)	D,E (14)	F (8)	42	4
2	A,C (15)	D,E (14)	B,F (15)	-	44	3

The assembly alternatives of Figure 1 represent two variants of an assembly process, which determine two alternative subgraphs: S1, which consists in performing tasks B, C, D and E (in the shown order); and S2, which consists in performing tasks C, D, E and B (in that order). A diagramming scheme called S-Graph, proposed by Capacho and Pastor [3], is used here to represent all alternative subgraphs in a unique precedence graph, as shown in Figure 2.

To solve the ASALBP, Capacho and Pastor [3] developed an integer linear programming mathematical model (hereafter referred to as the preliminary model, M1), which decides on both the assembly subgraphs and the line balancing. The

task-workstation assignment variables in M1 are defined for each total assembly precedence graph (i.e., global route) obtained by combining the alternative subgraphs of each available subassembly contained in the S-Graph. Therefore, each of the global routes considers the whole set of tasks required to assemble a product. Figure 3 shows two global routes (GR1 and GR2) for the example shown in Figure 1. The computational experiment carried out with M1 shows that optimal solutions can be obtained in a reasonable amount of time for small problems.

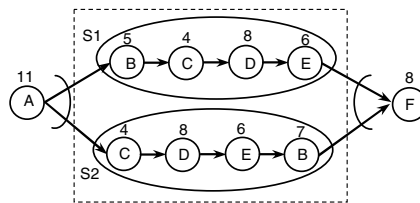


Fig. 2. Precedence S-Graph

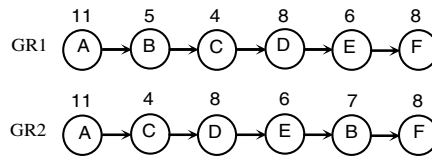


Fig. 3. Alternative global routes

To address more general problems, the definition of the ASALBP is extended in this paper: it is possible to consider alternative assembly processes that involve different and independent sets of tasks. Additionally, a new integer linear programming mathematical model (hereafter referred to as the enhanced model, M2) has been developed to formalize and solve this extended version of the ASALBP.

3 The Extended ASALBP Definition

As mentioned above, the former ASALBP definition considers the possibility of assembly alternatives in such a way that all alternatives of a particular sub-assembly (one of which must be selected) involve the same tasks. In practice, however, industrial problems can involve alternative assembly processes that consider different and mutually exclusive sets of tasks. For instance, consider the toy-manufacturing example given in Das and Nagendra [4], in which a large number of products are assembled from molded plastic parts or from metal stamping. In this example, the tasks are performed only if the assembly process they belong to is selected. Figure 4 shows the S-Graph for an example with three assembly processes involving 3 different and independent sets of tasks: S1 with tasks B1 to B3; S2 with tasks C1 to C4; and S3 with tasks D1 to D3.

The new ASALBP definition considers alternative assembly precedence subgraphs that can involve either the same or different sets of tasks. In these two cases, task processing times and/or precedence relations can be dependent on the selected

subgraph. As in the former definition, it is assumed that assembly alternatives do not overlap each other; thus, each alternative of each subassembly is represented by a unique and independent precedence subgraph. This new ASALBP definition enables more realistic problems to be addressed and on the other hand, relaxes the SALBP hypothesis which states that all tasks must be processed only once.

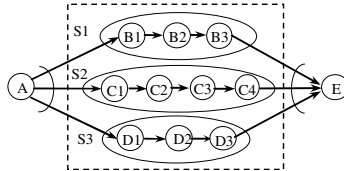


Fig. 4. S-Graph with alternative assembly processes

4 Enhanced Model for the Alternative Subgraphs Assembly Line Balancing Problem (M2)

Before presenting model M2, some aspects concerning assignment variables and precedence relations are discussed.

4.1 Modeling Assumptions

In model M1, each overall assembly alternative referred to as a global route was determined by a precedence graph. As a result, there were as many global assembly routes as combinations of alternative subgraphs for each available subassembly contained in the S-Graph. Consider, for example, the S-graph in Figure 5, which involves 9 assembly tasks. In this example, 9 global routes need to be considered as follows. GR1: A-S1-E-S4-I (i.e. A-B-C-D-E-F1-F2-F3-I), GR2: A-S1-E-S5-I (i.e. A-B-C-D-E-G1-G2-G3-I), GR3: A-S1-E-S6-I, GR4: A-S2-E-S4-I, GR5: A-S2-E-S5-I, GR6: A-S2-E-S6-I, GR7: A-S3-E-S4-I, GR8: A-S3-E-S5-I and GR9: A-S3-E-S6-I.

Observe that tasks that do not admit processing alternatives are also involved in all global routes. This implies that a large number of task-workstation assignment variables have to be defined for even a small number of routes. In the proposed model, alternative subgraphs (referred to as partial routes) are considered only for those tasks that admit processing alternatives and a unique route (known as a base route) is considered for those tasks without processing alternatives. Task-workstation

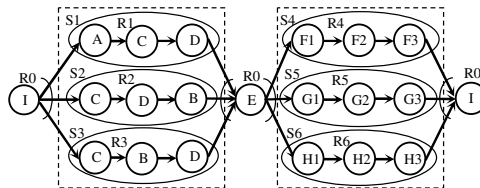


Fig. 5. S-Graph for an example with 9 tasks

assignment variables are thus defined only for the partial routes (alternative subgraphs), thereby considerably reducing the size of the model to be solved. For the example in Figure 5, tasks A, E and I have only one possible processing alternative. Therefore, such tasks are processed according to the base route (referred to as R0). Partial routes R1, R2 and R3 represent the processing alternatives involving tasks B, C and D; R4 is the processing alternative that involves tasks F1 to F3; R5 involves G1 to G3 and R6 involves H1 to H3. For each subset of routes that are alternative to each other (e.g., R4, R5 and R6), one partial route must be selected. Observe that only 7 partial routes are involved; the difference between the number of global and partial routes is even greater because in M1 the number of global routes increases exponentially with the number of partial routes.

For global routes, the immediate predecessors of a task for each route are fixed and the precedence constraints can be easily established. However, this is not the case for partial routes. The difficulty arises because an immediate predecessor or the task itself may have alternatives and only one of these is to be selected. Therefore, all possible immediate predecessors of a task must be considered.

In order to account for all possible precedence relations implied when considering partial routes and to facilitate their formalization, tasks can be divided into two categories: fixed, which are those without alternatives (i.e., those processed throughout the base route); and mobile, which are those that form a part of alternative routes. As can be seen in Figure 6, tasks A, F and G are fixed whereas tasks B, C, D and E are mobile because they are involved in alternative assembly routes: R1 and R2 involving tasks B and C, and R3 and R4 involving tasks D and E. A fictitious task with nil processing time, α , is used in the S-Graph to represent precedence relations that involve mobile tasks with predecessors that are also mobile but affected by different sets of alternative routes. This case is represented in Figure 6 by the mobile tasks D and E, whose predecessors C and B are also mobile tasks affected by different routes.

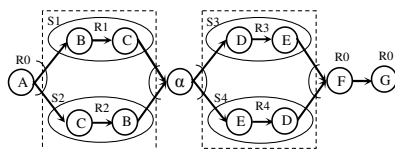


Fig. 6. Fixed and mobile tasks - precedence relations

Table 2 shows the 5 basic cases of task-predecessor relations, formalized below, which are based on the example in Figure 6.

Table 2. Task-predecessor relation typology

Cases		<i>i</i>	<i>p</i>
1	Task <i>i</i> fixed and its predecessor <i>p</i> fixed.	G	F
2	Task <i>i</i> fixed and its predecessor <i>p</i> mobile.	F	E,D
3	Task <i>i</i> mobile and its predecessor <i>p</i> fixed.	B	A
4	Task <i>i</i> mobile and its predecessor <i>p</i> mobile, with <i>i</i> and <i>p</i> in the same route.	C	B
5	Task <i>i</i> mobile and its predecessor <i>p</i> mobile, with <i>i</i> and <i>p</i> in different route.	D	C,B

4.2 Model M2

Parameters:

- n Number of tasks ($i = 1, \dots, n$).
- nr Number of routes ($r = 0, \dots, nr$).
- nsr Number of different sets of routes (subgraphs) such that the routes within a set are alternatives to each other ($q=1, \dots, nsr$). For instance, in the example of Figure 6 there are 2 such subsets ($nsr=2$), one containing routes R1 and R2, and one containing routes R3 and R4.
- m_{min}, m_{max} Lower and upper bounds on the number of stations.
- R_i Set of all routes through which task i can be processed ($i = 1, \dots, n$).
- ct Cycle time.
- t_{ir} Duration of task i when processed through route r ($i = 1, \dots, n; r \in R_i$).
- TR_r Set of tasks that are affected by route r .
- P_{ir} Set of the possible immediate predecessors of task i , if task i is processed through route r ($i = 1, \dots, n; r \in R_i$).
- PT_i Set of all possible immediate predecessors of task i ($PT_i = \bigcup_{r \in R_i} P_{ir}$).
- E_{ir}, L_{ir} Earliest and latest station that task i can be assigned to, if task i is processed through route r ($i = 1, \dots, n; r \in R_i$).
- SCR_q Subset q of routes that are alternative among one another ($q=1, \dots, nsr$). For the example in Figure 6, there are two of such subsets: SCR_1 involving R1 and R2 and SCR_2 involving R3 and R4.

Decision variables:

- $x_{ijr} \in \{0,1\}$ = 1 if task i is assigned to workstation j and processed through route r ($i=1, \dots, n; \forall r \in R_i; \forall j \in [E_{ir}, L_{ir}]$).
- $y_j \in \{0,1\}$ = 1 if there is any task assigned to workstation j ($j=m_{min}+1, \dots, m_{max}$).
- $ar_r \in \{0,1\}$ = 1 if there is any task processed through route r ($r = 1, \dots, nr$).

Mathematical Model for the ASALBP-1: to minimize the number of workstations given ct

$$\text{Minimize } Z = \sum_{j=m_{min}+1}^{m_{max}} j \cdot y_j \quad (1)$$

$$\sum_{j=E_{i0}}^{L_{i0}} x_{ij0} = 1 \quad \forall i \mid i \in TR_0 \quad (2)$$

$$\sum_{\forall r \in R_i} \sum_{j=E_{ir}}^{L_{ir}} x_{ijr} = \sum_{\forall r \in R_i} ar_r \quad \forall i \mid i \notin TR_0 \quad (3)$$

$$\sum_{r=0}^{nr} \sum_{\forall i \mid (r \in R_i) \wedge (j \in [E_{ir}, L_{ir}])} t_{ir} \cdot x_{ijr} \leq ct \quad j = 1, \dots, m_{min} \quad (4)$$

$$\sum_{r=0}^{nr} \sum_{\forall i \mid (r \in R_i) \wedge (j \in [E_{ir}, L_{ir}])} t_{ir} \cdot x_{ijr} \leq ct \cdot y_j \quad j = m_{min} + 1, \dots, m_{max} \quad (5)$$

$$\sum_{j=E_{p0}}^{L_{p0}} j \cdot x_{pj0} \leq \sum_{j=E_{i0}}^{L_{i0}} j \cdot x_{ij0} \quad \forall i \in TR_0, \forall p \in PT_i \mid p \in TR_0 . \tag{6}$$

$$\sum_{\forall s \in R_p} \sum_{j=E_{ps}}^{L_{ps}} j \cdot x_{pjs} \leq \sum_{j=E_{i0}}^{L_{i0}} j \cdot x_{ij0} \quad \forall i \in TR_0, \forall p \in PT_i \mid p \notin TR_0 . \tag{7}$$

$$\sum_{j=E_{p0}}^{L_{p0}} j \cdot x_{pj0} \leq \sum_{\forall r \in R_i} \sum_{j=E_{ir}}^{L_{ir}} j \cdot x_{ijr} + m_{\max} \cdot (1 - \sum_{\forall r \in R_i} ar_r) \quad \forall i \notin TR_0, \forall p \in PT_i \mid p \in TR_0 . \tag{8}$$

$$\sum_{j=E_{pr}}^{L_{pr}} j \cdot x_{pjr} \leq \sum_{j=E_{ir}}^{L_{ir}} j \cdot x_{ijr} \quad \forall i \notin TR_0, \forall r \in R_i, \forall p \in P_{ir} \mid [p \notin TR_0 \wedge r \in R_p] . \tag{9}$$

$$\sum_{\forall s \in R_p} \sum_{j=E_{ps}}^{L_{ps}} j \cdot x_{pjs} \leq \sum_{\forall r \in R_i} \sum_{j=E_{ir}}^{L_{ir}} j \cdot x_{ijr} \quad \forall i \notin TR_0, \forall p \in PT_i \mid [p \notin TR_0 \wedge (R_i \cap R_p = \emptyset)] . \tag{10}$$

$$\sum_{r \in SCR_q} ar_r = 1 \quad q = 1, \dots, nsr . \tag{11}$$

$$\sum_{\forall i \in TR_r} \sum_{j=E_{ir}}^{L_{ir}} x_{ijr} \leq ar_r \cdot |TR_r| \quad r = 1, \dots, nr . \tag{12}$$

The objective function (1) minimizes the number of workstations for a given upper bound on the cycle time. The constraints are: (2) and (3), which ensure that all tasks belonging to a selected route are assigned to one and only one workstation, and otherwise tasks are not assigned; (4) and (5) ensure that the total processing time assigned to workstation j does not exceed the cycle time; (6) to (10) are the precedence constraints, which guarantee that no task is assigned to an earlier workstation than an immediate predecessor; (11) are the route uniqueness constraints that ensure that one and only one route for each subassembly is selected from among the possible routes; and (12) guarantees that tasks belonging to a particular precedence subgraph are assigned to the same route.

The mathematical formulation for ASALBP-2, considering the extended definition, can be easily obtained by modifying model M2 developed to solve ASALBP-1, in which the objective function to be considered optimizes the cycle time ct instead of the number of workstations, which is a given parameter.

5 Computational Experiment

To solve problems involving alternative processes with different sets of tasks, model M1 was easily adapted by properly defining the global routes. To compare models M1 and M2 and evaluate their performance regarding industrial-sized problems, a computational experiment was carried out. The test-problem instances were designed using some of the benchmark data sets available at the webpage www.assembly-line-balancing.de for assembly line balancing research (i.e., the problems of Bowman, Mansor, Buxey, Gunther, Kilbrid, Hann, Warnecke, Tonge and Arc with 8, 11, 29, 35, 53, 58, 70 and 111 tasks respectively). The problems were adapted by incorporating a number of assembly alternatives (between 2 and 14) and using 3 or 4 different cycle-time values. When alternative assembly processes involving different tasks were considered, new sets of tasks were also added to the original problems. For example, 9 new tasks were added to Hann’s problem to contemplate 4 new assembly processes

involving 2, 3, 2 and 2 tasks respectively. A total of 82 test-problem instances were defined and solved with both models using the optimization software ILOG CPLEX 9.0 on a PC Pentium 4, CPU 2.88 GHz with 512 Mb of RAM. Table 3 presents the data and results for some of these problems, including the name of the problem, the number of tasks n , the cycle time ct , and the number of global routes for model M1 and partial routes for model M2.

Table 3. Results of optimally solving ASALBP instances

Problem	n	ct	No. of routes		Constraints		Variables		Solving Time		% of Improv.
			Global	Partial	M1	M2	M1	M2	M1	M2	
Bowman	8	20	18	9	366	77	1744	888	0.53	0.03	94,3
Mansor	11	62	12	8	288	74	804	547	0.63	0.09	85,7
Mansor	11	62	15	9	352	78	1002	614	2.10	0.16	91,0
Buxey	29	54	12	8	861	147	3850	2581	61547	92.03	99,9
Buxey	29	54	6	6	444	134	1936	1941	18485	0.86	100
Gunther	35	41	32	11	2633	205	25806	8911	89558	14805	83,5
Gunther	40	81	60	13	4287	189	28824	6276	467	0.31	99,9
Kilbrid	45	56	12	8	1383	204	10840	7247	213	1.41	99,3
Kilbrid	45	69	24	10	2505	217	17312	7241	830	1.06	99,9
Hann	53	4676	18	9	2424	238	4780	2403	114	0.13	99,9
Hann	58	2004	24	10	3400	263	19516	8157	8356	3.48	100
Hann	62	2806	36	12	5210	280	22340	7471	19785	249	98,7
Warnecke	58	111	2	3	368	235	3186	4754	7200	638	91,1
Warnecke	58	111	4	5	648	253	6318	7888	17709	1410	92,0
Tonge	70	185	8	7	1428	342	21356	18702	259200	80122	69,1
Average percentage of improvement of M2 over M1											93,6

Table 3 shows that model M2 outperformed model M1 in all cases. Furthermore, M2 achieved around 94 % of average improvement; reaching a 100% in nearly half of the problems solved. On the other hand, the number of variables and constraints was significantly reduced (as intended) and the solving time was considerably smaller than with the preliminary model M1. Therefore, optimal solutions could be obtained and guaranteed in a reasonable amount of time for problem instances involving up to 70 tasks and 40 assembly alternatives (i.e., 14 partial routes).

For the problems for which no optimal solution was guaranteed, models M1 and M2 were solved with the computing time restricted to 1800 seconds (a realistic time window in an industrial environment). Table 4 shows the results obtained with model M2; it presents no data for M1 because the established time limit was exceeded without any solution being provided for these problems. Observe that model M2 obtained solutions with one workstation deviation from the benchmark optimum. Furthermore, the known optimal solution was found for Gunther’s and Tonge’s problems (marked in Table 4 with an asterisk).

Table 4. Problems solved within a 1800-seconds time window

Problem name	No. of tasks	Cycle time	No. of routes		No. of workstations	
			Global	Partial	Obtained	Optimal
Gunther*	35	41	32	11	14	14
Warnecke	58	111	2	3	15	14
Tonge	70	320	8	7	12	11
Tonge*	70	220	8	7	17	17
Arc2	111	17067	4	5	10	9

6 Conclusions and Further Research

In this paper an extended definition of the ASALBP has been presented. More realistic problems can be addressed with this new definition since it allows alternative assembly processes to be considered that can involve either the same or different and mutually exclusive sets of tasks. Therefore, the hypothesis that states that tasks must be processed only once is relaxed because some tasks are not processed if the alternative they belong to is not selected.

Additionally, a mathematical programming model has been presented, which solves the ASALBP considering the extended definition. With this new model, the problems are solved in a significantly shorter time than with the preliminary model adapted to the new definition; moreover, medium-sized problems are solved in a very reasonable amount of time. For bigger problems, good feasible solutions were obtained in 1800 seconds, a realistic time window in an industrial environment. Nevertheless, since the solving time increases exponentially with the number of tasks and assembly alternatives, further research is needed in order to develop heuristics to solve the ASALBP regarding the extended definition presented and formalized in this paper.

References

1. Baybars, I. A survey of exact algorithms for the simple assembly line balancing problem. *Management Science*, 32, 909-932 (1986).
2. Becker, C. and Scholl, A. A survey on problems and methods in generalized assembly line balancing. *Eur. J. of Op. Res.*, 168, 694-715 (2006).
3. Capacho, L. and Pastor, R. ASALBP: the Alternative Subgraphs Assembly Line Balancing Problem. *Technical Report IOC-DT-P-2005-5*. UPC. Spain. Jan. (2005).
4. Das, S. and Nagendra, P. Selection of routes in a flexible manufacturing facility. *Int. Journal of Production Economics*, 48, 237-247 (1997).
5. Gen, M., Tsujimura, Y. and Li, Y. Fuzzy assembly line balancing using genetic algorithms. *Comp. & Ind. Eng.*, 31, 631-634 (1996).
6. Kim, Y.K., Kim, Y.J. and Kim, Y. Genetic algorithms for assembly line balancing with various objectives. *Comp. and Ind. Eng.*, 30, No.3, pp. 397-409 (1996).
7. Lapierre, S.D. and Ruiz, A.B. Balancing assembly lines: an industrial case study. *J. of the Operational Research Society*, 55, 559-597 (2004).
8. Miltenburg, J. Balancing and scheduling mixed-model U-shaped production lines. *International Journal of Flexible Manufacturing Systems*, 14, 119-151. (2002).
9. Pastor, R., Andres, C., Duran, A. and Perez, M. Tabu search algorithms for an industrial multi-product, multi-objective assembly line balancing problem, with reduction of task dispersion. *J. Op. Res. Society*, 53, 1317-1323 (2002).
10. Scholl, A. and Becker, C. State-of-the-art exact and heuristic solution procedures for simple assembly line balancing. *European J. of Op. Res.*, 168, 666-693 (2006).
11. Scholl, A. and Klein, R. SALOME: A bidirectional branch and bound procedure for assembly line balancing. *INFORMS Journal on Comp.*, 9, 319-334 (1997).
12. Suresh, G. and Sahu, S. Stochastic assembly line balancing using simulated annealing. *Int. Journal of Production Research* 32, 1801-1810 (1994).

Satisfying Constraints for Locating Export Containers in Port Container Terminals

Kap Hwan Kim and Jong-Sool Lee

Department of Industrial Engineering, Pusan National University,
Changjeon-dong, Kumjeong-ku, Pusan 609-735, Korea
{kapkim, jong-sool}@pusan.ac.kr

Abstract. To allocate spaces to outbound containers, the constraint satisfaction technique was applied. Space allocation is pre-assigning spaces for arriving ships so that loading operations can be performed efficiently. The constraints, which are used to maximize the efficiency of yard trucks and transfer cranes, were collected from a real container terminal and formulated in the form of constraint. Numerical experiments were conducted to evaluate the performance of the developed algorithm.

1 Introduction

Operations in port container terminals consist of the discharging operation (during which containers are unloaded from ships), the loading operation (during which containers are loaded onto ships), the delivery operation (during which inbound containers are transferred from the marshalling yard to outside trucks), and the receiving operation (during which outbound containers are transferred from outside trucks to the marshalling yard). The discharging operation and the loading operation are together called the “ship operation.” For sake of optimal customer service, the turnaround time of container-ships must be minimized by increasing the speed of the ship operation, and the turnaround time of outside trucks must be shortened as much as possible. Figure 1 shows a container yard.

In container terminals, the loading operation for outbound containers is carefully pre-planned by load planners. For the load planning, the responsible container-ship agent usually transfers a load profile (an outline of a load plan) to the terminal operating company several days before a ship’s arrival. In the load profile, each slot (cell) is assigned a container group, which is identified by type (full or empty), port of destination, and the size of container to be stowed onto. Because a cell of a ship can be filled with any container within its assigned group, the handling work in the marshalling yard can be facilitated by optimally sequencing outbound containers for the loading operation. In sequencing the containers, load planners usually attempt to minimize the handling of quay cranes and the yard equipment. The output of this decision-making process is called the “load sequence list.” To find an efficient load sequence, outbound containers must be laid out in the optimal location. The main focus of this paper is to suggest a method of pre-allocating storage space for arriving containers so that maximum efficiency is achieved in the loading operation.

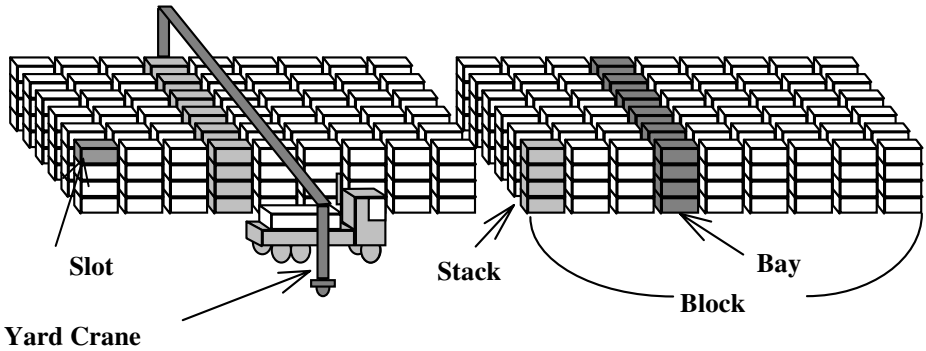


Fig. 1. An illustration of a container yard

In order to obtain an efficient load sequence, the following must be considered during the space planning process. Figure 2 shows a containership's cross-sectional view, which is called a ship-bay. The figure shows cells into which the containers of two groups (defined by size and port of destination) are assigned. A widely accepted principle for space planning is that yard-bays assigned to a containership should be located near the berthing position of the corresponding ship. In addition, there may be other principles of space planning that depend on the type of yard equipment. One such example is that containers of different groups should not be mixed in the same yard-bay. This principle is valid only for the indirect transfer (combined) system of yard-side equipment (yard crane or straddle carrier) and prime movers. During the loading operations of containers, containers of the same group are likely to be loaded onto cells located close together, as illustrated in Figure 2, and thus, the containers are usually loaded consecutively. Therefore, for the case of the indirect transfer system, the travel distance of the yard-side equipment can be reduced by placing containers of the same group in the same yard-bay. In addition, there are many practical rules that yard planners use for allocating spaces to different groups of containers. The rules will be discussed further in the following sections.

There have been many related studies regarding container terminals. Taleb-Ibrahimi [1] analyzed the space-allocation problem with a constant or cyclic space requirement for stacking containers. Kim and Kim [2] formulated a quadratic mixed-integer-programming model for the dynamic space-allocation problem, but they did not suggest an efficient algorithm for the mathematical model. Kim and Kim [3] addressed the space allocation problem for inbound container. Kim and Kim [4] discussed the factors that affect the efficiency of the loading operation of outbound containers. Kim et al. [5] suggested a method for determining storage locations for outbound containers so that the number of rehandles during the loading operation is minimized. Cao and Uebe [6] suggested a transportation model with a non-linear constraint for assigning available space to space requirements. However, they did not consider the dynamic aspect of container flows over the time horizon. Kozan [7] proposed a network model to describe the flow of containers in port container terminals. The model attempted to determine flows of different types of containers in a way that minimizes the total handling cost. Roll and Rosenblatt [8] suggested a grouped

storage policy that is based on a concept similar to the space-allocation problem in container terminals. They applied the group storage strategy as a storage policy for warehouses. Tsang [9] described the constraint satisfaction technique in detail. Zhang *et al.* [10] discussed the storage space allocation problem in the storage yards of container terminals. They decomposed the space allocation problem into two levels: the subproblem in the first level attempts to balance workloads among different yard blocks, while the second subproblem minimizes the total transportation distance for moving containers between blocks and vessel berthing locations. Kim and Park [11] proposed a multicommodity minimal cost flow problem model for the space allocation problem. A subgradient optimization technique was applied to solve the problem.

All the previous studies assumed that the objective function is clearly defined, and that feasible solutions can be easily obtained. However, in container terminals, there are many complicated constraints to be satisfied, and so, finding a feasible solution itself is a difficult problem. This is why CSP technique is applied to the space allocation problem in container terminals.

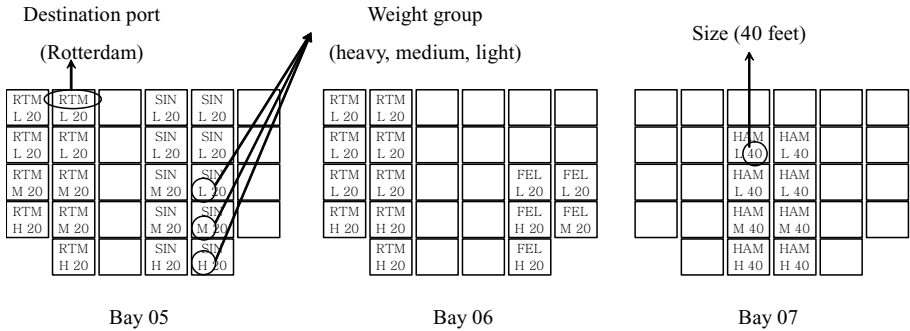


Fig. 2. An example of a stowage plan of a container-ship

2 Space Allocation Problem for Export Containers

It is assumed that the allocation of space is performed periodically. The length of the allocation period may be one day, 12 hours, or 6 hours, depending on the level of uncertainty and the time of the computation. Each period is called a “stage” in the decision-making process of space allocation.

The level of inventory in containers that arrive at a container yard follows a similar pattern. Arriving containers are classified into container groups, each of which is a collection of containers of the same length, vessel, and destination port. It is also assumed that containers of different groups are not stacked in the same yard-bay. The space must be pre-allocated for each group of containers that will arrive during the next stage. However, if decisions for the next stage are made without considering future changes in the yard, it may be impossible to find a feasible solution for the succeeding stages. Thus, an investigation must be performed on the effects of the decisions for the next stage on those for the subsequent stages. In this study, the investigation is performed by the CSP technique.

By using the forecasted arrival of containers, space requirements are estimated for each group of containers that will arrive at the yard in the next stage and the subsequent stages. A container group that requires an allocation of space is called an SDU (Space Demand Unit). The amount of space needed by each SDU is expressed in the unit of one yard-bay for 20-ft containers and two yard-bays for 40-ft containers. Based on the expected arrival of containers, SDUs for the next stage and the subsequent stages, all of which represent the demand side of the space allocation, must be specified.

Next, the supply side of space allocation must be considered. A container yard for outbound containers is usually divided into several blocks, each of which consists of 20 to 30 yard-bays. Each yard-bay consists of 20 to 30 stacks in a straddle carrier system and 6 to 8 stacks in a yard crane system. Space can be allocated in units of stacks or yard-bays, depending on the type of handling equipment and the space-allocation strategy used. In this study, a yard-bay is considered to be the unit of space allocation (SAU). Figure 3 shows the conceptual representation of the space allocation for this study. The space allocation assigns one or two available SAUs to an SDU. No SAU can be assigned to more than one SDU.

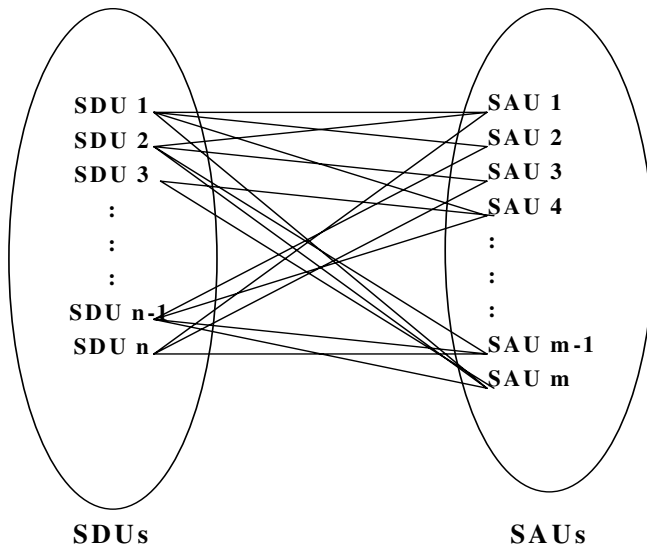


Fig. 3. Matching SDUs with SAUs

One of the difficulties of the space allocation problem is that the quality of the allocation decisions can be evaluated only when the loading operation is performed. However, the efficiency of the loading operation is dependent on the load sequencing of the outbound containers as well as the allocations. Because load sequencing is another complicated decision-making problem and there are many complicated constraints to be satisfied for the space allocation, the optimization is not a practical approach for the space allocation problem.

This paper discusses how the technique for the constraint satisfaction problem (CSP) can be applied to the space allocation problem. The following explains the constraints that were collected from an actual container terminal. Most of the constraints are related to rules that have been used for a long time by yard planners for efficient loading operations in container terminals.

(Constraint 1) The distance between the berth that a vessel arrives at and the location of a block where the outbound containers of the vessel are stacked must be less than a specified maximum limit. This constraint is necessary to reduce the travel cost of yard trucks between the apron and the yard.

(Constraint 2) The maximum distance between blocks where containers for one vessel are located must be less than a specified value. This constraint is necessary to reduce the travel distance of transfer cranes.

(Constraint 3) Containers for one vessel must be stacked in the blocks that are located in the same row of the yard. This constraint is necessary because transfer cranes can travel more easily in the lengthwise direction of blocks than in the widthwise direction of blocks.

(Constraint 4) A block's space cannot be allocated to the receiving operation of a vessel when the loading operation of another vessel is scheduled at the same time at the same block. This constraint is necessary to prevent the congestion of transfer cranes in the same block.

(Constraint 5) The number of vessels onto which containers stacked in a block will be loaded cannot exceed a specified limit (NV_{max}). This constraint has the effect of simultaneously restricting both the maximum number of blocks to be allocated to a vessel and the minimum number of containers to be stacked for a vessel.

(Constraint 6) The number of blocks, in which the containers to be loaded onto the same vessel are stacked, cannot exceed a specified limit (NB_{max}). When the containers for one vessel are scattered over too many blocks, the travel distance of the transfer cranes may be excessive.

(Constraint 7) A 40-ft container requires two consecutive 20-ft yard-bays.

In addition to the above constraints, other constraints can be additionally considered without significantly modifying the search algorithm.

3 Application of the CSP Techniques to the Allocation of Spaces

Figure 4 shows the structure of the program developed for the space allocation problem. The system consists of an interface layer, a constraint specification layer, and a search layer. In the interface layer, variables and their domains are specified. In the space allocation problem, variables correspond to SDUs, while the domain of an SDU corresponds to SAUs that can be allocated to the SDU. In the constraint specification layer, constraints, which are expressed in the form of equations, are specified. Various program modules are already provided and can be easily used only by specifying the values of parameters.

The following describes the search procedure in this study:

Step 1: Define the variables and domain, which is a set of values that the variable can take, of each variable.

- Step 2: If there remains no more variable, then stop. Otherwise, select the next variable.
- Step 3: Select the next value. Assign the selected value to the variable. If all the variables are assigned values, then stop. Otherwise, go to Step 4.
- Step 4: Reduce the problem. In this step, values of the remaining variables, which do not satisfy at least one constraint, will be removed from the domains of the variables. Check if there is a variable whose domain becomes empty. If yes, then go to Step 5. If no, then go to Step 2.
- Step 5: Check if there remains any value to assign for the current variable. If yes, then go to Step 3. If no, then let the current variable be the previous variable (backtracking) and go to Step 3.

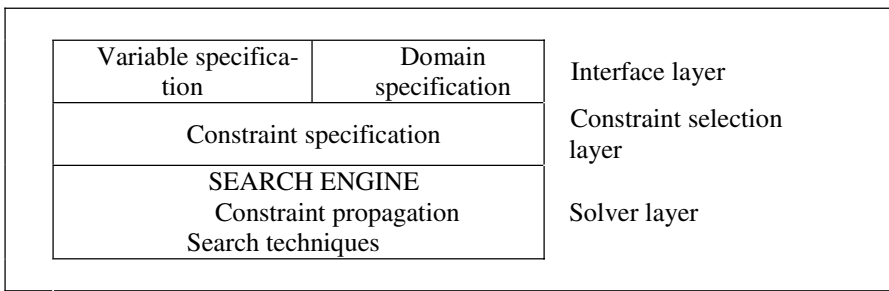


Fig. 4. The structure of the program developed for the space allocation

4 Numerical Experiment

Numerical experiment was conducted to test the performance of the search algorithm and find the best search strategies.

4.1 Input Data for the Numerical Experiment

The algorithm developed in this study was applied to solve a real space allocation problem of a large container terminal (PECT: Pusan Eastern Container Terminal) in Pusan. Various search strategies were tested to evaluate the speed of the search algorithm. The strategies used were variable-ordering rules, value-ordering rules, and constraint-ordering rules. A problem with two stages, 84 variables (SDUs) which approximately equal to 15 (vessels) \times 2 (sizes) \times 3 (destination ports), and 600 values (SAUs) of each variable, which corresponds to the number of bays in the yard, was solved. The data set came from a practical case with 15 vessels and 24 blocks, and 4 berths. The problems in the experiment considered all seven constraints mentioned in section 2. Parameters for constraints 5 and 6 were set as follows: $NV_{\max} = 3$ and $NB_{\max} = 3$.

4.2 Experiment to Evaluate Various Variable-Ordering Strategies

The following three criteria were used for ordering variables.

- (1) Stage of SDU: The SDUs of earlier stages have higher priorities than those of later stages.
- (2) Size of containers of SDU: The SDUs of 40-ft containers have higher priorities than those of 20-ft containers.
- (3) Vessel of SDU: The SDUs of vessels arriving at a terminal are prioritized based on chronological order.

By combining the three different criteria, three variable-ordering rules were constructed as follows:

(Rule 1) Sequence SDUs according to the stage criteria.

(Rule 2) Sequence SDUs according to the size criteria.

(Rule 3) Sequence SDUs according to the stage criteria first, and followed by the vessel criteria.

(Rule 4) Sequence SDUs according to the stage criteria first, the vessel criteria second, and the size criteria third.

SDUs with the same values of sequencing criteria are sequenced in a random order.

The values in the domains are sequenced in the order of increasing bay ID. The problem was solved for ten initial distributions of containers. Results in Table 1 show the CPU time to find a feasible solution for ten problems.

Through a statistical test, three null hypotheses that the computational time by rule 4 is not greater than that by each of the other three rules were rejected under the significance level of 1%. The results of the hypothetical test imply that using rule 4 results in a shorter computation time, compared to the other rules.

Table 1. The computational time for various variable-ordering rules (in seconds)

Initial distribution	Rule 1	Rule 2	Rule 3	Rule 4
1	654	596	556	497
2	670	602	570	521
3	643	579	554	504
4	665	588	586	487
5	663	607	564	479
6	657	577	565	518
7	655	589	573	510
8	647	590	546	496
9	660	610	577	488
10	662	588	573	505
Average	658	593	566	501

4.3 Experiment to Evaluate Two Value-Ordering Strategies

An experiment was performed on different sequences of values in the domains. Rule 3 was used as the variable-ordering rule. Two rules for value-ordering were compared with each other. The first rule is to sequence SAUs in the alphabetical order of the bay ID, and this rule will be called the “bay ID rule.” The second rule is to give higher priorities to SAUs that are located in the blocks nearer to the berthing location of the vessel corresponding to the SDU. The second rule will be called the “closest-to-berth rule.”

As in the case of the first experiment, ten problems with different initial distributions of stacked containers were solved. The results of the numerical experiment are shown in Table 2.

Table 2. The computational time for two value-ordering rules (in seconds)

Initial distribution of containers	Bay ID rule	Closest-to-berth rule
1	556	537
2	570	546
3	554	532
4	586	566
5	564	563
6	565	535
7	573	549
8	546	532
9	577	545
10	573	558
Average	566	543

By a statistical test, it was concluded that the closest-to-berth rule outperforms the bay ID rule in computational time under the significance level of 1%.

4.4 Experiment to Evaluate Various Constraint-Ordering Strategies

There are many constraints that solutions of the space allocation problem must satisfy. The propagation sequence of constraints during the search process is expected significantly affect the computational time, which will be tested in this subsection. Constraints 3, 5, 6, and 7 were considered. It was assumed that $NV_{\max} = 3$ and $NB_{\max} = 3$. Rule 3 and the bay ID rule were used as the variable-ordering rule and the value-ordering rule, respectively.

For each sequence of constraints, ten problems with different initial distributions of containers were solved. Table 3 shows the average computational time of the ten test problems for different sequence of constraints. The table shows that the sequence of constraints significantly affects the computational time and that constraint 5 should be propagated first during the search process.

Table 3. The computational time for different sequences of constraints

Seq.	Sequence of constraints	Search time (in s)	Seq.	Sequence of constraints	Search time (in s)
1	5→7→6→3	465	13	3→6→7→5	564
2	5→6→3→7	472	14	3→5→6→7	567
3	5→7→3→6	477	15	3→7→5→6	567
4	5→3→7→6	497	16	3→7→6→5	570
5	6→5→3→7	498	17	3→5→7→6	584
6	5→3→6→7	501	18	3→6→5→7	592
7	5→6→7→3	503	19	7→5→3→6	598
8	6→7→5→3	509	20	7→3→6→5	607
9	6→7→3→5	512	21	7→3→5→6	612
10	6→3→5→7	513	22	7→5→6→3	629
11	6→3→7→5	520	23	7→6→5→3	630
12	6→5→7→3	532	24	7→6→3→5	632

5 Conclusions

The constraint satisfaction problem (CSP) technique was applied to a space allocation problem for outbound containers. A program that realized the CSP concept was developed for the space allocation. Constraints for the space allocation problem were introduced.

Using real data collected from the Pusan Eastern Container Terminal, Korea, numerical experiments were conducted to evaluate the performance of the developed algorithm. Various variable-ordering rules were compared with each other in terms of their computational time. The results showed that sequencing space demand (requirement) units by the stage criteria first, the vessel criteria second, and the size criteria third results in the shortest computational time. It was also shown that the value-ordering rule significantly affects the computational time. Lastly, various sequences of constraint propagation during the search process were compared with each other. It was also shown that the sequence of constraints significantly affects the computational time.

Acknowledgments

This work was supported by the Regional Research Centers Program (Research Center for Logistics Information Technology), granted by the Korean Ministry of Education & Human Resources Development.

References

1. Taleb-Ibrahimi, M., Castilho, B., and Daganzo, C. F.: Storage Space vs Handling Work in Container Terminals. *Transportation Research* 127B (1) (1993) 13-32
2. Kim, K. H. and Kim, D. Y.: Group Storage Methods at Container Port Terminals. *The Material Handling Engineering Division 75th Anniversary Commemorative Volume ASME MH-Vol.2* (1994) 15-20
3. Kim, K. H. and Kim, H. B.: Segregating Space Allocation Models for Container Inventories in Port Container Terminals. *International Journal of Production Economics* 59 (1999) 415-423
4. Kim, K. H. and Kim, K. Y.: An Optimal Routing Algorithm for a Transfer Crane in Port Container Terminals. *Transportation Science* 33(1) (1999) 17-33
5. Kim, K. H., Park, Y. M., and Ryu, K. R.: Deriving Decision Rules to Locate Export Containers in Container Yard. *European Journal of Operational Research* 124 (2000) 89-101
6. Cao, B. and Uebe, G.: Solving Transportation Problems with Nonlinear Side Constraints with Tabu Search. *Computers Ops Res.* 22(6) (1995) 593-603
7. Kozan, E.: Optimizing Container Transfers at Multimodal Terminals. *Mathematical and Computer Modelling* 31 (2000) 235-243
8. Roll, Y. and Rosenblatt, M. J.: Random versus Grouped Storage Policies and Their Effect on Warehouse Capacity. *Material Flow* 1 (1983) 199-205
9. Tsang, E.: *Foundations of Constraint Satisfaction*. Academic Press Limited, UK (1993)
10. Zhang, C., Liu, J., Wan, Y.-W., Murty, K. G, and Linn, R. J.: Storage Space Allocation in Container Terminals. *Transportation Research*. 37B (2003) 883-903
11. Kim, K. H. and Park, K. T.: Dynamic Space Allocation for Temporary Storage. *International Journal of Systems Science* 34 (2003) 11-20.

A Price Discrimination Modeling Using Geometric Programming

Seyed J. Sadjadi¹ and M. Ziaee²

Iran University of Science and Technology
sjsadjadi@iust.ac.ir

Abstract. This paper presents a price discrimination model which determines the product's selling price and marketing expenditure in two markets. We assume production as a function of price and marketing cost in two states. The cost of production is also assumed to be a function of production in both markets. We propose a Multi Objective Decision Making (MODM) method called Lexicograph Approach (LA) in order to find an efficient solution between two objectives for each market. Geometric Programming is used to solve the resulted model. In our GP implementation, we use a transformed dual problem in order to reduce the model to an optimization of a nonlinear concave function subject to some linear constraints and solved the resulted model using a simple grid line search.

Keywords: Mathematical Programming, Production and Operations Management, Economics.

1 Introduction

One of the most important issues on having a fair price discrimination strategy is on choosing a right model. Many traditional discrimination models assume production as a function of price. The production function often is in a form of linear or quadratic. On the other hand, economists normally study the cost of production as a function of production with similar linear or quadratic pattern. The production and cost functions considered in this paper have an exponential form of price, marketing expenditure and production, respectively. This type of modeling have been widely used in the literature [4, 7, 8, 10]. They consider demand as a function of price and marketing expenditure and assume that when demand increases production will be less costly. Sadjadi et. al. [10] study the effects of integrated production and marketing decisions in a profit maximizing firm. Their model formulation is to determine price, marketing expenditure, demand or production volume, and lot size for a single product with stable demand when economies of scale are given. Hoon and Cerry [5] present a method for optimal inventory policies for an economic order quantity with decreasing cost functions. Lee [7] considers the same demand function to determine order quantity and selling price. In their implementation, they use a previous [8] model formulation, with an adaptation of geometric programming(GP), to determine

the global solution of the model. The primary assumption of this paper is to determine price and marketing strategy in two markets. Suppose a fortune 500 company which markets its product in North America and plans to penetrate into a new market with the same advertisement expenditure. Therefore, we have a new market with a new price strategy. However, due to a big advertisement strategy, the company has already introduced its brand and does not need for any additional advertisement. Therefore, we assume that demands in both markets are functions of different prices but unique marketing cost. The objective function of our modeling is to maximize the profits for both markets. However, the primary objective is to maximize the profitability in the first market and the profit maximization in the second market is our secondary objective. Therefore, we have to maximize the profitability for two objective functions. The resulted MODM problem is solved using Lexicograph Approach where we maximize the profit for the first state and then we optimize the second market's profit keeping the optimal profitability of the first market as a constraint. The resulted problem in each stage is in posynomial GP problem [3]. We use GP method to find the optimal solution of the resulting model. This paper is organized as follows. We first present the problem statement. Next, the problem statement is presented in MODM form. GP method is used to find the optimal solution of the problem formulation. Throughout the paper, we use a numerical example in order to show the implementation of the algorithm. Finally, concluding remarks are presented at the end to summarize the contribution of the work.

2 Problem Statement

Consider a single product where demand is affected by selling price. Let P_i , α_i , M and γ_i be the selling price per unit, price elasticity to demand, marketing expenditure per unit and marketing expenditure elasticity to demand in the market $i = 1, 2$, respectively. For both markets we assume,

$$D_i = k_i P_i^{-\alpha_i} M^{\gamma_i}, \quad i = 1, 2. \tag{1}$$

where productions (D_i , $i = 1, 2$) are defined as a function of price per unit (P_i) and marketing expenditure (M) with $\alpha_i > 1$, $0 < \gamma_i < 1$, $i = 1, 2$. The scaling constants k_i represent other related factors and the assumption $\alpha_i > 1$, $i = 1, 2$ implies that D_1 and D_2 increase at a diminishing rate as P_1 and P_2 decrease. This type of relationship is widely used in the literature [7, 8, 9, 10]. Besides, (1) can be easily estimated by applying linear regression to the logarithm of the function. Let c_1 and c_2 be the production cost per unit for state one and two, respectively. We assume that the unit production cost (c_1) and (c_2) can be discounted with β_1 and β_2 , respectively. Therefore we have,

$$c_1 = u_1 D_1^{-\beta_1}, \quad c_2 = u_2 D_2^{-\beta_2}, \tag{2}$$

where D_1 and D_2 are production lot size(units), u_1 and u_2 are the scaling constants for unit production cost in state one and two, respectively . The exponent

β_1 and β_2 represent lot size elasticity of production unit cost with $0 < \beta_1, \beta_2 < 1$ which are almost the same as price elasticity α and we suggest a small value for it, say $\beta_1, \beta_2 = 0.01$. We will also explain that the algorithm we use imposes some other limitations for all the parameters in our model.

2.1 The Proposed Model

In this section, we present our proposed production lot sizing and marketing model ($\pi_i, i = 1, 2$) based on the explained assumptions. For the first we are interested in maximizing the profit $\pi(P_i, M)$ simultaneously in order to determine the prices and marketing expenditure for the planning horizon as follows,

$$\begin{aligned} \max \quad & \pi_i(P_i, M) = \text{Revenue in Market } i \\ & \text{-Production cost in Market } i \\ & \text{- Marketing expenditure in Market } i \\ & = P_i D_i - C_i D_i - M D_i, \quad i = 1, 2. \end{aligned} \tag{3}$$

As we can observe in (3), we are interested in maximization of π_1 and π_2 . Let π_1^* be the optimal solution for π_1 . Using an auxiliary variable t we have,

$$\begin{aligned} \max \quad & P_2 D_2 - C_2 D_2 - M D_2, \\ \text{subject to} \quad & \pi_1 = P_1 D_1 - C_1 D_1 - M D_1 \geq t \pi_1^*. \end{aligned} \tag{4}$$

In order to solve (4), we need to have the optimal solution π_1^* . The optimal solution for π_1 is obtained as follows,

$$\max z = P_1 D_1 - C_1 D_1 - M D_1. \tag{5}$$

Problem (5) is in Geometric Programming which can be easily formulated in posynomial form. Since there are two variables and three terms associated with (5) the degree of difficulty is equal to $3-(2+1)=\text{zero}$ [3]. Therefore we have,

$$\begin{aligned} \max \quad & z = \pi_1 \text{ or } \min z^{-1} \\ \text{subject to} \quad & P_1 D_1 - C_1 D_1 - M D_1 \geq z \end{aligned} \tag{6}$$

or

$$\begin{aligned} \min \quad & z^{-1} \\ \text{subject to} \quad & k_1 P_1^{1-\alpha_1} M^{\gamma_1} - u_1 k_1^{1-\beta_1} P_1^{\alpha_1(\beta_1-1)} M^{\gamma_1(1-\beta_1)} - k_1 P_1^{-\alpha_1} M^{\gamma_1+1} \geq z. \end{aligned} \tag{7}$$

Therefore we have,

$$\begin{aligned} \min \quad & z^{-1} \\ \text{subject to} \quad & u_1 k_1^{-\beta_1} P_1^{\alpha_1 \beta_1 - 1} M^{-\beta_1 \gamma_1} + P_1^{-1} M + k_1^{-1} P_1^{\alpha_1 - 1} M^{-\gamma_1} z \leq 1. \end{aligned} \tag{8}$$

Problem (8) is in posynomial form and can be solved using its dual problem formulation as follows,

$$\begin{aligned}
 d(\pi_1) &= \max f(w) = \left[\frac{1}{w_0}\right]^{w_0} \left[\frac{u_1 k_1^{-\beta_1} \lambda}{w_1}\right]^{w_1} \left[\frac{\lambda}{w_2}\right]^{w_2} \left[\frac{k_1^{-1} \lambda}{w_3}\right]^{w_3} \\
 \text{subject to} \quad & w_0 = 1 \\
 & -w_0 + w_3 = 0 \\
 & (\alpha_1 \beta_1 - 1)w_1 - w_2 + (\alpha_1 - 1)w_3 = 0 \\
 & -\beta_1 \gamma_1 w_1 + w_2 - \gamma_1 w_3 = 0.
 \end{aligned} \tag{9}$$

Thus,

$$\begin{aligned}
 w_1 &= (\gamma_1 + 1 - \alpha_1)/(\alpha_1 \beta_1 - \beta_1 \gamma_1 - 1), \quad w_3 = 1 \\
 w_2 &= (\beta_1 \gamma_1 - \gamma_1)/(\alpha_1 \beta_1 - \beta_1 \gamma_1 - 1), \quad \lambda = (\alpha_1 \beta_1 - \alpha_1)/(\alpha_1 \beta_1 - \beta_1 \gamma_1 - 1)
 \end{aligned} \tag{10}$$

Using $w_i, i = 0, \dots, 3$ from (10), one can determine the optimal solution π_1^* from (9) and solve (4) as follows,

$$\begin{aligned}
 \min \quad & \pi_1^{*-1} t^{-1} + z_2^{-1} \\
 \text{subject to} \quad & P_1 D_1 - C_1 D_1 - M D_1 \geq t \pi_1^* \\
 & P_2 D_2 - C_2 D_2 - M D_2 \geq z_2.
 \end{aligned} \tag{11}$$

or

$$\begin{aligned}
 \min \quad & \pi_1^{*-1} t^{-1} + z_2^{-1} \\
 \text{subject to} \quad & u_1 k_1^{-\beta_1} P_1^{\alpha_1 \beta_1 - 1} M^{-\beta_1 \gamma_1} + P_1^{-1} M + \pi_1^* k_1^{-1} P_1^{\alpha_1 - 1} M^{-\gamma_1} t \leq 1, \\
 & u_2 k_2^{-\beta_2} P_2^{\alpha_2 \beta_2 - 1} M^{-\beta_2 \gamma_2} + P_2^{-1} M + k_2^{-1} P_2^{\alpha_2 - 1} M^{-\gamma_2} z_2 \leq 1.
 \end{aligned} \tag{12}$$

Problem (12) is a minimization of a nonlinear posynomial objective function subject to two posynomial constraints. Since there are eight terms and five variables, the degree of difficulty is $8 - (5 + 1) = 2$. Let

$$\begin{aligned}
 \delta_{01} &= \pi_1^{*-1} t^{-1}, & \delta_{02} &= z_2^{-1} \\
 \delta_{11} &= u_1 k_1^{-\beta_1} P_1^{\alpha_1 \beta_1 - 1} M^{-\beta_1 \gamma_1} & \delta_{12} &= P_1^{-1} M \\
 \delta_{13} &= \pi_1^* k_1^{-1} P_1^{\alpha_1 - 1} M^{-\gamma_1} t & \delta_{21} &= u_2 k_2^{-\beta_2} P_2^{\alpha_2 \beta_2 - 1} M^{-\beta_2 \gamma_2} \\
 \delta_{22} &= P_2^{-1} M & \delta_{23} &= k_2^{-1} P_2^{\alpha_2 - 1} M^{-\gamma_2} z_2.
 \end{aligned} \tag{13}$$

Let w_{ij} be the dual variables associated with δ_{ij} for $i = 0, 1, 2$ and $j = 1, 2, 3$, respectively. Therefore we have,

$$d(\pi_2) = \max f(w) = \left[\frac{\pi_1^{-1} \lambda_0}{w_{01}} \right]^{w_{01}} \left[\frac{\lambda_0}{w_{02}} \right]^{w_{02}} \left[\frac{w_1 k_1^{-\beta_1} \lambda_1}{w_{11}} \right]^{w_{11}} \left[\frac{\lambda_1}{w_{12}} \right]^{w_{12}} \left[\frac{\pi_1 k_1^{-1} \lambda_1}{w_{13}} \right]^{w_{13}} \times$$

$$\left[\frac{w_2 k_2^{-\beta_2} \lambda_2}{w_{21}} \right]^{w_{21}} \left[\frac{\lambda_2}{w_{22}} \right]^{w_{22}} \left[\frac{k_2^{-1} \lambda_2}{w_{23}} \right]^{w_{23}}$$

subject to

$$\begin{aligned} -w_{01} + w_{13} &= 0 \\ -w_{02} + w_{23} &= 0 \\ (\alpha_1 \beta_1 - 1)w_{11} - w_{12} + (\alpha_1 - 1)w_{13} &= 0, \\ (\alpha_2 \beta_2 - 1)w_{21} - w_{22} + (\alpha_2 - 1)w_{23} &= 0, \\ -\beta_1 \gamma_1 w_{11} + w_{12} - \gamma_1 w_{13} - \beta_2 \gamma_2 w_{21} + w_{22} - \gamma_2 w_{23} &= 0, \\ \lambda_0 &= w_{01} + w_{02} = 1, \\ \lambda_1 &= w_{11} + w_{12} + w_{13}, \\ \lambda_2 &= w_{21} + w_{22} + w_{23}. \end{aligned} \tag{14}$$

We rewrite the linear equations of (14) in terms of two dual variables, w_{01} and w_{11} . Therefore we have,

$$\begin{aligned} w_{13} &= w_{01}, \quad w_{23} = 1 - w_{01}, \quad w_{12} = (\alpha_1 \beta_1 - 1)w_{11} + (\alpha_1 - 1)w_{01}, \quad w_{02} = 1 - w_{01}, \\ w_{21} &= ((\alpha_1 \beta_1 - \beta_1 \gamma_1 - 1)w_{11} + (\alpha_1 - \gamma_1 - \alpha_2 + \gamma_2)w_{01} + (\alpha_2 - \gamma_2 - 1)) \\ &\quad / (\beta_2 \gamma_2 + 1 - \alpha_2 \beta_2), \\ w_{22} &= (\alpha_2 - 1)(1 - w_{01}) + (\alpha_2 \beta_2 - 1)w_{21}. \end{aligned} \tag{15}$$

As we can observe, the linear constraints in (14) can be converted into (15) where there are only two unknowns. Therefore, we may use a simple grid search to find the optimal solution. Note that in order to have a feasible solution in (14) the following must hold,

$$\begin{aligned} l_1 &= (\alpha_1 - 1)/(1 - \alpha_1 \beta_1), \quad l_2 = (\alpha_1 - \gamma_1 - \alpha_2 + \gamma_2)/(\beta_1 \gamma_1 + 1 - \alpha_1 \beta_1) \\ l_3 &= (\alpha_2 - \gamma_2 - 1)/(\beta_1 \gamma_1 + 1 - \alpha_1 \beta_1), \\ l_4 &= [(\alpha_1 - 1 - \gamma_1)(1 - \alpha_2 \beta_2) + \gamma_2(1 - \beta_2)]/[(\beta_1 \gamma_1 + 1 - \alpha_1 \beta_1)(1 - \alpha_2 \beta_2)], \\ l_5 &= [\gamma_2(1 - \beta_2)]/[(\beta_1 \gamma_1 + 1 - \alpha_1 \beta_1)(1 - \alpha_2 \beta_2)], \\ \max [0, l_4 w_{01} + l_5] &< w_{11} < \min [l_1 w_{01}, l_2 w_{01} + l_3] \end{aligned} \tag{16}$$

Once the optimal values of dual variables are obtained, we may choose the relationships to determine the optimal primal variables, P_1^* , P_2^* and M^* as follows,

$$\delta_{ij}^* = w_{ij}^* / \lambda_i^* \quad \lambda_i^* \neq 0, \quad i = 1, 2, \quad j = 1, 2, 3. \tag{17}$$

Once we determine the optimal values of δ_{ij}^* , we may use (13) in order to find the optimal values of P_1^* , P_2^* and M^* . Next section, we demonstrate the implementation of the proposed method of this paper using a numerical example.

2.2 Numerical Example

In this section we present numerical experience of the implementation of the proposed method. Suppose we have,

$$\alpha_1 = 1.5, \alpha_2 = 2.0, \beta_1 = 0.01, \beta_2 = 0.02, \gamma_1 = .1,$$

$$\gamma_2 = .2, u_1 = 4, u_2 = 5, k_1 = k_2 = 1000000.$$

This example is solved using the procedure explained in this section. The procedure first finds the optimal solution, π_1^* . In the second phase we find the optimal solution of π_2^* subject to $\pi_1 \geq \pi_1^*$. The optimal weights are calculated to be

$$w_{01}^* = 0.99, w_{02}^* = 0.01, w_{11}^* = 0.3997, w_{12}^* = 0.1013, w_{13}^* = 0.99, w_{21}^* = 0.0103,$$

$$w_{22}^* = 9.8 \times 10^{-5}, w_{23}^* = 0.01, \lambda_1^* = 1.491, \lambda_2^* = 0.0204,$$

and the optimal solution is summarized as follows,

$$P_1^* = \$13.5162, P_2^* = \$8.1668, M^* = \$0.9187, \pi_1^* = 1.9741 \times 10^5,$$

$$\pi_2^* = 4.6014 \times 10^4, \pi^* = \pi_1^* + \pi_2^* = 2.4343 \times 10^5.$$

Example (1) demonstrates the implementation of the proposed method of this paper. Since the resulted dual model can be simply run when different input parameters are changed, one can practically use this method in order to determine the optimal production and marketing expenditure in two states. Although, a sensitivity analysis has been extensively used by others [6, 7, 8] for single objective model, we believe that the argument would not be necessarily used as an applicable tool for multi-objective model presented in this paper.

3 Conclusion

In this paper, we have presented a price discrimination model which determines the product's selling price and marketing expenditure in two markets. We have assumed that the production is a function of price and marketing expenditure in two states and the cost of production is also assumed to be a function of production in both markets. A Multi Objective Decision Making (MODM) method called Lexicograph Approach (LA) has been proposed to find an efficient solution between two objectives for each market. Geometric Programming has been used to solve the resulted model. In our GP implementation, we have used a transformed dual problem in order to reduce the model to an optimization of a nonlinear concave function subject to some linear constraints. The resulted model has been solved using a simple grid line search.

References

1. Beightler, C. S. and D. T. Phillips: Applied geometric programming. New York: Wiley, (1976)
2. Dembo, R. S.: Sensitivity analysis in geometric programming. *Journal of Optimization Theory and Applications*, **37** (1982) 1–21
3. Duffin, R. J., Peterson, E. L., & Zener: Geometric programming—Theory and application. New York: John Wiley & Sons, (1967)
4. Freeland, J. R.: Coordination strategies for production and marketing in a functionally decentralized firm. *AIIE Transactions*, **12** (1982) 126–132
5. Hoon, J. and M. K. Cerry: Optimal inventory policies for an economic order quantity model with decreasing cost functions. *European Journal of Operational Research*, **165** (2005) 108–126
6. Kim, D. and Lee, W. J.: Optimal joint pricing and lot sizing with fixed and variable capacity. *European Journal of Operational Research*, **109** (1998) 212–227
7. Lee, & W. J.: Determining selling price and order quantity by geometric programming, Optimal solution, bounds and sensitivity. *Decision Sciences*, **24** (1993) 76–87
8. Lee, W. J., and D. Kim: Optimal and heuristic decision strategies for integrated production and marketing planning. *Decision Sciences*, **24** (1993) 1203–1213
9. Lilien, G. L., Kotler, P. Moorthy, and K. S.: Marketing models. Englewood Cliffs, NJ: Prentice Hall, (1992)
10. Sadjadi S. J., M. Orougee, and M. B. Aryanezhad: Optimal Production and Marketing Planning. *Computational Optimization and Applications*, **30(2)** (2005)
11. Starr M. K.: Managing production and operations. Englewood Cliffs, NJ: Prentice Hall, (1989)

Hybrid Evolutionary Algorithms for the Rectilinear Steiner Tree Problem Using Fitness Estimation*

Byounggak Yang

Department of Industrial Engineering, Kyungwon University,
San 65 Bockjung-dong , Sujung-gu, Seongnam-si, Kyunggi-do, Korea
byang@kyungwon.ac.kr

Abstract. The rectilinear Steiner tree problem (RSTP) is to find a minimum-length rectilinear interconnection of a set of terminals in the plane. A key performance measure of the algorithm for the RSTP is the reduction rate that is achieved by the difference between the objective value of the RSTP and that of the minimum spanning tree without Steiner points. We introduced four evolutionary algorithm based upon fitness estimation and hybrid operator. Experimental results show that the quality of solution is improved by the hybrid operator and the calculation time is reduced by the fitness estimation. The best evolutionary algorithm is better than the previously proposed other heuristics. The solution of evolutionary algorithm is 99.4% of the optimal solution.

1 Introduction

Given a set V of n terminals in the plane, the rectilinear Steiner tree problem(RSTP) in the rectilinear plane is to find a shortest network, a Steiner minimum tree, interconnecting S . The points in S are called Steiner points. We should find the optimal number of Steiner points and their location on rectilinear plane. It is well-known that the solution to RSTP will be the minimal spanning tree (MST) on some set of points $V \cup S$. Let $MST(V)$ be the cost of MST on set of terminals V . Then the reduction rate $R(V) = \{MST(V \cup S) - MST(V)\} / MST(V)$ is used as performance measure for algorithms on the Steiner tree problem.

The RSTP is known to be NP-complete [8]. Polynomial-time algorithm for the optimal solution is unlikely to be known [7]. Warme, Winter and Zachariasen [17] presented exact algorithm for RSTP and showed the average reduction rate is about 11.5% for the optimal solution. Lee, Bose and Hwang [14] presented algorithm similar to the Prim algorithm for the Minimum Spanning tree. Beasley[3] presented standard instances for Steiner tree problem and heuristics algorithms for the RSTP[4]. The best heuristics of RSTP are the Batched Iterated 1-Steiner (BIIS) of Kahng [13] and the Edge-based heuristic of Borah [5] with average reduction rate of 11%. The genetic algorithm for Euclidian Steiner tree problem was presented by Hesser et al. [10] and Barreiros[2], but they didn't introduce algorithm for RSTP. Julstrom[12] introduced three kinds of genetic algorithm for RSTP. In his research, the genetic algorithm

* This research was supported by the Kyungwon University Research Fund in 2005.

based on lists of edges was relatively better than the other genetic algorithm. But, reduction rate of genetic algorithms were worse than the other heuristics. Sock and Ahn[15] used genetic algorithm to solve the degree constrained minimum spanning tree. Yang[19] introduced an evolutionary algorithm for RSTP; the 11.0% of reduction rate of the evolutionary algorithm by Yang was almost near that of optimal solution. But, the computation load for evaluating the cost function was quite heavy. Main motivation of this research is to develop another evolutionary algorithm to find near optimal solution within reasonable calculation time.

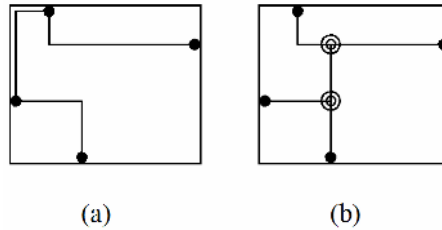


Fig. 1. A minimum rectilinear spanning tree (a) and a minimum rectilinear Steiner tree (b)

Hanan showed that for any instance, an optimal RST exists in which every Steiner point lies at the intersection of two orthogonal lines that contain terminals [9]. Hanan's theorem implies that a graph G called the Hanan grid graph is guaranteed to contain an optimal RST. Hanan's grid graph is constructed as follows: draw a horizontal and vertical line through each terminal. The vertices in graph correspond to the intersections of the lines. For 3 terminals RST problem, we can find optimal Steiner point easily. Let (x_i, y_i) be the coordinates of the given terminal T_i ; France [6] proved that the Steiner point S is located at (x_m, y_m) , where x_m and y_m are the medians of $\{x_i\}$ and $\{y_i\}$.

2 Evolutionary Algorithm

Evolutionary algorithms are based on models of organic evolution. They model the collective learning process within a population of individuals. The starting population is initialized by randomization or some heuristics method, and evolves toward successively better regions of search space by means of randomized process of recombination, mutation and selection. Each individual is evaluated as the fitness value, and the selection process favors those individuals of higher quality to reproduce more often than worse individuals. The recombination mechanism allows for mixing of parental information while passing it to their dependents, and mutation introduces innovation into the population. Although simplistic from a biologist's viewpoint, these algorithms are sufficiently complex to provide robust and powerful adaptive search mechanisms [1]. In this research, some Steiner points in Hanan's grid may become an individual, and the individual is evaluated by minimum spanning tree with terminals and Steiner points in the individual.

For most evolutionary algorithm, a large number of fitness evaluations are needed and fitness evaluation is not trivial. To reduce the evaluating complexity, we can employ fitness approximation. One of the fitness approximations is evolutionary approximation that fitness evaluations can be shared by estimating the fitness value of the child from the fitness value of their parents [11].

2.1 General Evolutionary Algorithm

By the Hanan’s theorem, the optimal Steiner points should be on the Hanan’s grid. An individual are introduced as represented the location of Steiner point on Hanan’s grid. Each vertical line and horizontal line is named as increasing number. Most left vertical line is the first vertical line, and most bottom horizontal line is the first horizontal line. Let v_i be the index of vertical line on Steiner point S_i and h_i be the index of horizontal line on Steiner point S_i then each individual has multiple (v_i, h_i) which is the location of i -th Steiner point S_i on Hanan’s grid. And (x_i, y_i) is the real location on plane for Steiner point S_i , and is calculated from (v_i, h_i) . In this research, individual is an assembly of (v_i, h_i) of a non-fixed number of Steiner points and number of Steiner points as follows ; $\{m, (v_1, h_1), (v_2, h_2), \dots, (v_m, h_m)\}$ where m is the number of Steiner points.

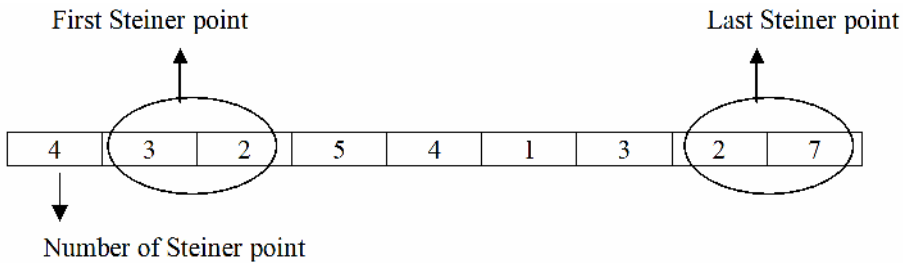


Fig. 2. The individual for evolutionary algorithm

Each individual has different length depending on its number of Steiner points, and represents one Steiner Tree. The fitness of the individual corresponds to the length of the MST that can be constructed by using the original terminals and the Steiner points in individual by Prim algorithm.

The recombination or crossover operator exchanges a part of an individual between two individual. We choose some Steiner points in each individual and exchange with other Steiner points in the other individual. We have two parents to crossover as $P_1=\{3, (1,3), (2,4), (3,2)\}$ and $P_2=\{2, (6,4), (7,8)\}$. By the random number, we choose the number of exchanging which is less than the number of Steiner points in both parents. Then by the random number, some Steiner points are choose and exchanged. In this case, we choose the 2 as exchanging number. First and third Steiner points are selected in P_1 , and two Steiner points are selected in P_2 . Therefore we have two children by performing crossover as follows: $O_1=\{3, (6,4), (2,4), (7,8)\}$ and $O_2=\{2, (1,3), (3,2)\}$.

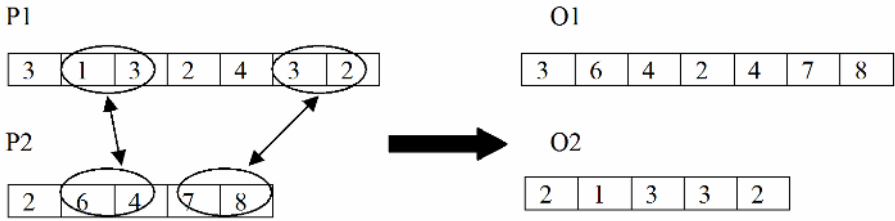


Fig. 3. The crossover operator

The mutation operator changes the locations of some selected Steiner points. Let (v_i, h_i) be the selected location of Steiner point. Then we can have new location $(v_i + v, h_i + h)$ where v and h are some random integer number. We use the tournament selection. The k -tournament selection method select a single individual by choosing some number k of individuals randomly from the population and selecting the best individual from this group to survive to the next generation. The process is repeated as often as necessary to fill the new population. A common tournament size is $k=2$.

For the 3-terminal RST problem, the median of terminals is the optimal Steiner point. We assume that a median point of some adjacent 3-terminal has high probability to be survived in the optimal Steiner tree. For the convenience to find adjacent 3-terminal, we make a minimum spanning tree for the V and find every directly connected 3 terminals in the minimum spanning tree as like Fig. 4(b). And a Steiner point for these connected 3 terminals in minimum spanning tree is calculated. We make Steiner points pool as like Fig. 4(c) and use one of candidate Steiner point. We make initial population for the evolutionary algorithm as following procedure:

- Step 1: Make a minimum spanning tree (MST) for the terminals.
- Step 2: Choose every connected three terminals in MST. Let them be 3-neighbor terminals (3NT).
- Step 3: Make a Steiner point for each 3NT and add it in the Steiner points pool.
- Step 4: Choose some Steiner points from Steiner points pool by random function and make one individual. Repeat Step 4 until all individuals are decided.

We make an initial population with 200 individuals.

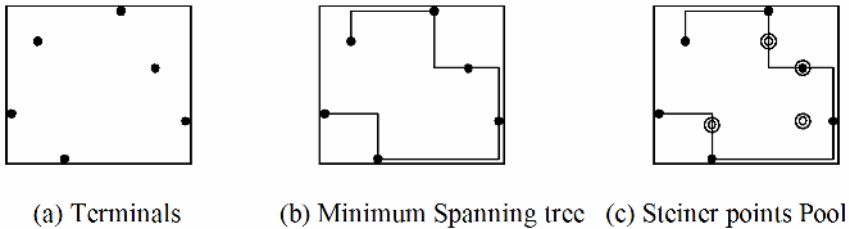


Fig. 4. A Steiner points pool from minimum spanning tree

2.2 Hybrid Operator

For searching new solution in evolutionary algorithm, the local search can be used. In this research, an insertion operator, a deletion operator and a moving operator are introduced. Mutation operator and crossover have the location of Steiner point moved to search a solution, but they don't search new Steiner point. We need to introduce a new Steiner point in current individual to enforce the variability of the solution. For some selected individual, randomly generated Steiner point in Hanan's grid is inserted in individual and increase the number of Steiner point. On the other hand, some Steiner points in tree are connected with only one or two other node. Those kinds of Steiner point in tree are obviously useless to reduce the tree cost. We introduce deletion operator to delete poor Steiner points which are connected less than 2 other node in Steiner tree. For the Steiner point connected with 3 other node, we know the optimal location of Steiner point as using the results of 3-terminal case. We trace the Steiner point (S) which is connected with exactly 3 other node(A,B,C) and calculate the optimal location(S*) of Steiner point for those 3 node(A,B,C) and move the location of S to S* as a moving operator.

2.3 Estimated Fitness Evaluation

We should solve the MST problem to evaluate each the individual. In Fig. 5(a), we have an evaluated parent individual. By some evolutionary operator as like mutation or crossover, some Steiner points are moved as like Fig. 5(b). As we know the tree structure of its parent, we can let the tree structure of child be that of parent. Then the tree cost of child is calculated simply by updating the cost of arcs which is directly connected to the moved Steiner points as like Fig. 5(c). If we solve the MST algorithm, then we get the MST as like Fig. 5(d) which has exact or minimum tree cost of child. Therefore the tree cost of Fig.5(c) is the estimation of child fitness. The difference between estimated value and exact value is increasing while the evaluation is re-estimated. Therefore we perform exact evaluation for all individuals to refresh the value of individuals if there is no improvement in solution for some given number of iterations. We use this kind of estimated evaluation to reduce the evaluation time.

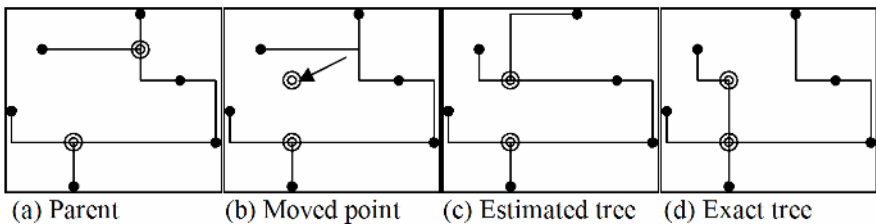


Fig. 5. The estimated evaluation for the child using the tree structure of parent

2.4 Alternative Algorithms

We introduced four evolutionary algorithms. First one is the general evolutionary algorithm (EV) without the estimated evaluation and the hybrid operator. Second one

is the hybrid evolutionary algorithm (EA_H) which uses the hybrid operator. Third one is the estimated evolutionary algorithm (EA_E) which uses the estimated evaluation. And last one is the hybrid estimated evolutionary algorithm (EA_HE) which uses the both estimated evaluation and hybrid operator.

3 Computational Experience

The computational study was made on Pentium IV processor. The evolutionary algorithm was programmed in Visual Studio. Problem instances are from OR-Library [3], 15 instances for each problem size 10, 20... 100, 250, 500, 1000. First test is to compare the reduction rate and calculation time of four evolutionary algorithms. In Fig. 6(a), the reduction rates of EA_H and EA_HE are better than those of EA and EA_E in every size of problems. Therefore it is better to use the hybrid operator for finding good solution. In Fig. 6(b), the calculation times of EA_E and EA_HE are smaller than those of EA and EA_H. The average calculation time of EA_HE was 24% of that of EA_H. Therefore the estimated evaluation is acceptable for reducing the calculation time. As considering the both of the reduction rate and the calculation time, the EA_HE is the best algorithm in these four evolutionary algorithms. Second test is to compare EA_HE and EA_H with other heuristics and optimal solutions. For the RST problem, Beasley has solved same instances as our ones [4]. The optimal solutions of same instances are shown by Warme[17]. The result of Borah's heuristic [5] and Kahng's one [13] and Julstrom's genetic algorithm [12] are shown in their papers. Those results are in Table1. In Table1, the evolutionary algorithms introduced in this paper give larger percentage reduction (about 11.1%) in most instances than the other heuristics. Table 2 shows the tree cost of solution and the % gap which is gap of

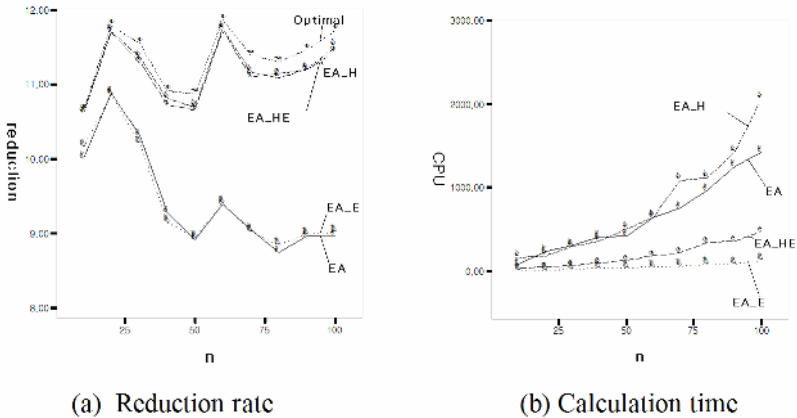


Fig. 6. The reduction rate and the calculation time of EA, EA_H, EA_E and EA_HE

tree cost between EA_HE, EA_H and optimal solution. The solution of evolutionary algorithm is 99.4% of the optimal solution.

In order to compare the evolutionary algorithm with optimal solutions, the 46 test problem were solved by the evolutionary algorithm. The 46 test problems were given by Soukup and Chow [16] and their optimal solution are shown in Warme's web site [18]. Table 3 shows the Steiner tree cost of optimal solutions, and those of the EA_H and EA_HE. The evolutionary algorithms found every optimal solution except only two instances which are the instances 43 and 44 in Table 3.

Table 1. The reduction rate of EA_HE, EA_H, Beasley's, Kahng's, Borah's and the optimal solution

<i>n</i>	EA_HE	EA_H	Beasley	Kahng	Borah	Julstrom	Optimal
10	10.625	10.611	9.947	10.36	10.33	*	10.656
20	11.711	11.721	10.590	10.44	10.4	*	11.798
30	11.306	11.356	10.250	*	*	*	11.552
40	10.724	10.807	9.956	*	*	*	10.913
50	10.655	10.705	9.522	10.71	10.71	9.167	10.867
60	11.736	11.755	10.146	*	*	*	11.862
70	11.099	11.163	9.779	*	*	2.8	11.387
80	11.142	11.081	9.831	*	*	*	11.301
90	11.186	11.203	10.128	*	*	4.06	11.457
100	11.445	11.501	10.139	10.89	10.84	0.72	11.720
250	11.113	11.090	9.964	10.88	10.88	*	11.646
500	10.706	10.706	9.879	*	10.94	*	11.631
1000	10.483	10.414	9.888	*	11.04	*	11.653
mean	11.072	11.086	10.001	10.656	10.734	4.187	11.419

* : They didn't show the results

Table 2. The tree cost of EA_HE, EA_H and the optimal solution

<i>n</i>	EA_HE	EA_H	Optimal	%gap of EA_HE	%gap of EA_H
10	2.21948194	2.219786293	2.218712547	0.034677468	0.048395033
20	3.296306033	3.295892153	3.293057227	0.098656247	0.086087987
30	4.12676294	4.124317633	4.11490786	0.28810074	0.228675189
40	4.9060314	4.901290273	4.895676353	0.211514118	0.114670979
50	5.375625147	5.372616413	5.36284096	0.238384594	0.182281246
60	5.71908182	5.717780633	5.710776713	0.145428671	0.122643913
70	6.266763493	6.262297327	6.246542193	0.323719898	0.252221675
80	6.76866624	6.773256053	6.756550107	0.179324258	0.247255573
90	7.12752516	7.126108713	7.105630147	0.308136124	0.288201978
100	7.543638527	7.53888388	7.520270667	0.310731635	0.247507226
250	11.68976045	11.69272585	11.61949869	0.604688336	0.630209288
500	16.36012215	16.34098039	16.19058313	1.047145854	0.928918076
1000	23.29998097	23.30472768	22.99475535	1.327370562	1.348013153
mean	8.053826636	8.051589484	8.002292458	0.393682962	0.363467794

Table 3. The comparison of the tree costs of Souckup and Chow's problems

Instances	n	EA_HE	EA_H	Optimal solution
1	5	1.87	1.87	1.87
2	6	1.64	1.64	1.64
3	7	2.36	2.36	2.36
4	8	2.54	2.54	2.54
5	6	2.26	2.26	2.26
6	12	2.42	2.42	2.42
7	12	2.48	2.48	2.48
8	12	2.36	2.36	2.36
9	7	1.64	1.64	1.64
10	6	1.77	1.77	1.77
11	6	1.44	1.44	1.44
12	9	1.8	1.8	1.8
13	9	1.5	1.5	1.5
14	12	2.6	2.6	2.6
15	14	1.48	1.48	1.48
16	3	1.6	1.6	1.6
17	10	2	2	2
18	62	4.04	4.04	4.04
19	14	1.88	1.88	1.88
20	3	1.12	1.12	1.12
21	5	1.92	1.92	1.92
22	4	0.63	0.63	0.63
23	4	0.65	0.65	0.65
24	4	0.3	0.3	0.3
25	3	0.23	0.23	0.23
26	3	0.15	0.15	0.15
27	4	1.33	1.33	1.33
28	4	0.24	0.24	0.24
29	3	2	2	2
30	12	1.1	1.1	1.1
31	14	2.59	2.59	2.59
32	19	3.13	3.13	3.13
33	18	2.68	2.68	2.68
34	19	2.44	2.43	2.41
35	18	1.51	1.51	1.51
36	4	0.9	0.9	0.9
37	8	0.9	0.9	0.9
38	14	1.66	1.66	1.66
39	14	1.66	1.66	1.66
40	10	1.55	1.55	1.55
41	20	2.24	2.24	2.24
42	15	1.53	1.53	1.53
43	16	2.57*	2.57*	2.55
44	17	2.52	2.54*	2.52
45	19	2.2	2.2	2.2
46	16	1.5	1.5	1.5

4 Conclusion

In this paper, we introduced a hybrid operator and an estimated evaluation. Four evolutionary algorithms for rectilinear Steiner tree problem are suggested. Computational

results showed that the hybrid operator improves the value of solution. The reduction rate of EA_HE and EA_H is about 11.1% which is similar to 11.4% of the optimal solutions. The estimated fitness evaluation reduces the calculation time by 76% without losing the quality of solution.

References

1. Bäck, Thomas : Evolutionary Algorithm in Theory and Practice, Oxford University Press (1996)
2. Barreiros, Jorge, : An Hierarchic Genetic Algorithm for Computing (near) Optimal Euclidean Stein Steiner Trees. *Workshop on Application of hybrid Evolutionary Algorithms to NP-Complete Problems*, Chicago, 2003
3. Beasley, J. E.: OR-Library: distributing test problems by electronic mail. *Journal of the Operational Research Society* 41 (1990) 1069-1072
4. Beasley, J. E: A heuristic for Euclidean and rectilinear Steiner problems. *European Journal of Operational Research* 58 (1992) 284-292
5. Borah, M., Owens, R. M. : An Edge-Based Heuristic for Steiner Routing. *IEEE Trans. on Computer Aided Design* 13(1994) 1563-1568
6. France, R.L.: A note on the optimum location of new machines in existing plant layouts, *J. Industrial Engineering* 14(1963) 57-59
7. Ganley, Joseph L.: Computing optimal rectilinear Steiner trees: A survey and experimental evaluation. *Discrete Applied Mathematics* 90 (1999) 161-171
8. Garey, M. R., Johnson, D. S.: The rectilinear Steiner tree problem is NP-complete. *SIAM Journal on Applied Mathematics* 32(1977) 826-834
9. Hanan, M.: On Steiner's problem with rectilinear distance. *SLAM Journal on Applied Mathematics* 14. (1966) 255-265
10. Hesser, J., Manner, R., Stucky, O. : Optimization of Steiner Trees using Genetic Algorithms, *Proceedings of the Third International Conference on Genetic Algorithm* (1989) 231-236
11. Jin, Y.: A Survey on fitness Approximation in Evolutionary Computation. *Journal of Soft Computing* 9(2005) 3-12
12. Julstrom, B.A. : Encoding Rectilinear Trees as Lists of Edges. *Proceedings of the 16th ACM Symposium on Applied Computing* (2001) 356-360
13. Kahng, A. B., Robins, B. : A New Class of Iterative Steiner Tree Heuristics with Good Performance. *IEEE Trans. on Computer Aided Design* 11(1992) 893-902
14. Lee, J. L., Bose, N. K., Hwang, F. K.: Use of Steiner's problem in suboptimal routing in rectilinear metric. *IEEE Transaction son Circuits and Systems* 23(1976) 470-476
15. Sock, S.M., Ahn, B.H. : A New Tree Representation for Evolutionary Algorithms. *Journal of the Korean Institute of Industrial Engineers* 31(2005) 10-19
16. Soukup, J., Chow, W.F.: Set of test problems for the minimum length connection networks. *ACM/SIGMAP Newsletter* 15(1973) 48-51
17. Warme, D.M., Winter, P., Zachariassen, M. : Exact Algorithms for Plane Steiner Tree Problems : A Computational Study In: D.Z. Du, J.M. Smith and J.H. Rubinstein (eds.): *Advances in Steiner Tree*, Kluser Academic Publishers (1998)
18. Warme, D.M. : [Http://www.group-w-inc.com/~warme/research](http://www.group-w-inc.com/~warme/research)
19. Yang, B.H. : An Evolution Algorithm for the Rectilinear Steiner Tree Problem. *Lecture Notes in Computer Science* 3483(2005) 241-249

Data Reduction for Instance-Based Learning Using Entropy-Based Partitioning

Seung-Hyun Son and Jae-Yearn Kim

Department of Industrial Engineering, Hanyang University,
17 Haengdang-Dong, Sungdong-Ku,
Seoul, 133-791, South Korea
shson@ihanyang.ac.kr, jyk@hanyang.ac.kr

Abstract. Instance-based learning methods such as the nearest neighbor classifier have proven to perform well in pattern classification in several fields. Despite their high classification accuracy, they suffer from a high storage requirement, computational cost, and sensitivity to noise. In this paper, we present a data reduction method for instance-based learning, based on entropy-based partitioning and representative instances. Experimental results show that the new algorithm achieves a high data reduction rate as well as classification accuracy.

1 Introduction

As competition among corporations intensifies and awareness of the importance of information grows, data mining methods that can extract useful information from large amounts of data are receiving increased interest. Information discovered through data mining can facilitate informed decision-making. Among data mining's several methods, classification techniques create models that distinguish data classes. A model is used to predict the class of objects whose class label is unknown. For example, a classification model may be built to categorize bank loan applications as either safe or risky, and used in customer confidence assessments by credit card companies, as well as in many marketing fields.

This paper presents a data reduction method for instance-based learning that is designed to improve classification accuracy through data reduction using entropy-based partitioning and representative instances. Also, through this method, the original data set is purged of irrelevant attributes and the number of instances is decreased.

Several data reduction methods for classification purposes have been proposed. Liu, Hussain, Tan, and Dash introduce a data reduction method that differentiates between attribute values [1], and Cano, Herrera, and Lozano introduce a combination of stratification and evolutionary algorithms [2]. Datta and Kibler introduced the prototype learner, which finds representative instances in each partition after dividing by the attribute value of each class [3]. They proposed a symbolic nearest mean classifier, which uses k -means clustering to group instances of the same class [4]. Wai Lam's prototype generation filtering (PGF)

algorithm operates by combining nearest instances and preferentially calculating the distance of each instance [5]. J.S. Sanchez's reduction by space partitioning (RSP) algorithms divide the training set into several subsets based on its diameter [6]. The diameter of a set is defined as the distance between its two farthest instances.

Most existing research methods use clustering techniques to create several partitions that maintain the homogeneity of the data [3, 4, 5, 6]. These clustering techniques calculate the distances of all instances repeatedly and create clusters. However, if the size of data increases, the computing time increases greatly. Also, these methods consider all attributes, including those that are irrelevant, which further increases computing time.

This paper presents an algorithm that accelerates partitioning as compared to existing methods and can remove irrelevant attributes. In addition, this new method can find the representative instances of each partition more quickly.

Section 2 of this paper introduces instance-based learning and measures of entropy and distance; Section 3 describes the procedure used by the proposed algorithm; Section 4 applies an example to explain the proposed algorithm; Section 5 presents experimental results; and finally Section 6 presents conclusions.

2 Preliminaries

This section introduces instance-based learning and measures for finding data partitioning and center instances.

2.1 Instance-Based Learning

The instance-based learning is a machine learning technique that has proven to be successful over a wide range of classification problems. The instance-based knowledge representation uses the instances themselves to represent what is learned, rather than inferring a rule set or decision tree and storing it instead.

The nearest neighbor algorithm is one of the most widely studied examples of instance-based learning methods [7, 8, 9]. This algorithm retains all of the training set and classifies unseen cases by finding the class labels of instances that are closest to them. It learns very quickly, because it only needs to read a training set without much further processing, and it generalizes accurately for many applications. Despite its high classification accuracy, however, it has a relatively high storage requirement and because it must search through all instances to classify unseen cases, it is slow to perform classification. Data reduction for instance-based learning can be used to obtain a reduced representation of the data, while minimizing the loss of information content.

2.2 Entropy Measure

Let S be a set consisting of s data instances. Suppose the class label attribute has m distinct values defining m distinct classes, C_i (for $i = 1, \dots, m$). Let s_i

be the number of instances of S in class C_i . The expected information needed to classify a given instance is given by

$$I(s_1, s_2, \dots, s_m) = - \sum_{i=1}^m p_i \log_2(p_i) , \tag{1}$$

where p_i is the probability that an arbitrary instance belongs to class C_i and is estimated by $\frac{s_i}{s}$ [10].

Let attribute A_1 have v distinct values, $\{a_1, a_2, \dots, a_v\}$. Attribute A_1 can be used to partition S into v subsets, $\{S_1, S_2, \dots, S_v\}$, where S_j contains those instances in S that have value a_j of A_1 . Let s_{ij} be the number of instances of class C_i in a subset S_j . The entropy based on the partitioning into subsets by A_1 , is given by

$$E(A_1) = \sum_{j=1}^v \frac{s_{1j} + \dots + s_{mj}}{s} I(s_{1j}, \dots, s_{mj}) . \tag{2}$$

The term $\frac{s_{1j} + \dots + s_{mj}}{s}$ acts as the weight of the j th subset and is the number of instances in the subset divided by the total number of instances in S . The smaller the entropy value is, the greater the purity of the subset partitions.

2.3 Distance Measure

The distance measure used to find the center instance in each partition is the Euclidean distance (ED) and is given by

$$ED(x, y) = \sqrt{\sum_{i=1}^a d(x_i, y_i)^2} , \tag{3}$$

where x and y are two instances, a is the number of attributes, and x_i refers to the i th attribute value, for instance x [10]. For numerical attributes, $d(x_i, y_i)$ is defined as their absolute difference (i.e., $|x_i - y_i|$). For categorical attributes, the distance between two values is typically given by

$$d(x_i, y_i) = 0 \quad \text{if } x_i = y_i, \text{ and } 1 \text{ otherwise} . \tag{4}$$

The center instance is based on the sum of the distances between instances in each partition, i.e., the center instance x^{ith} of each partition is given by

$$\min \left\{ \sum_{k=1}^n ED(x^{1st}, y^{kth}), \sum_{k=1}^n ED(x^{2nd}, y^{kth}), \dots, \sum_{k=1}^n ED(x^{nth}, y^{kth}) \right\} , \tag{5}$$

where n is the number of instances in each partition, x^{ith} and y^{kth} are the i th instance and the k th instance, respectively. The center instance is decided based on the least Euclidean distance measure. These formulas are used at the step that locates the center instance.

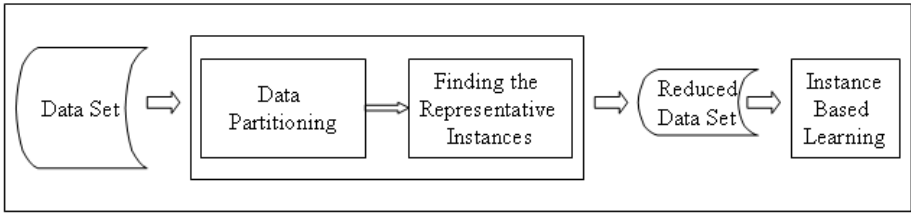


Fig. 1. Proposed algorithm process

3 Proposed Algorithm

The proposed algorithm consists of parts that seek the data partition and the representative instances. First, we calculate the entropy of each attribute. The data set is segmented preferentially via the attribute that has the smallest entropy. Using this method, data having homogeneity are gathered in the same partition, and each partition has the characteristics of the original data set. Second, we locate the representative instances using Euclidean distance. The representative instances consist of instances that represent the characteristics of each partition.

The procedure used by the proposed algorithm is as follows:

Steps 1–4: Data partitioning

- Step 1.* Calculate the entropy of all attributes using Equations (1) and (2). Select the attribute that has the lowest entropy. Partition the data set via the attribute's values.
- Step 2.* In each partition, calculate the entropy of the remaining attributes and redivide the partition via the attribute that has the smallest entropy value in each partition.
- Step 3.* This partitioning process continues until all partitions are pure, meaning all the class values are the same, or no further partitioning is possible.
- Step 4.* Several partition sets are composed.

Steps 5–8: Finding the representative instances

- Step 5.* Find the center instance of each partition. The center instance is determined by using Equations (3), (4), and (5). In this case, not all attributes are considered to find the center instance; attributes that are used once at each partition can be ignored because they have the same value in each partition set, and attributes that are not used at the data partitioning stage are regarded as irrelevant attributes and are thus purged.
- Step 6.* In each partition, find the k nearest instances to the center instance; k is proportional to the number of instances in each partition.
- Step 7.* In each partition, find the representative instances. The representative instances consist of the union of the center instance and the k nearest instances to the center instance.

#	A1	A2	A3	A4	A5	A6	Class
1	0	0	0	0	0	1	0
2	0	0	0	1	1	1	0
3	0	0	1	0	0	1	0
4	0	0	1	1	0	0	1
5	0	1	0	0	0	1	0
6	0	1	0	1	1	1	0
7	0	1	1	0	1	0	0
8	0	1	1	1	0	0	1
9	1	0	0	0	1	1	0
10	1	0	0	1	1	0	0
11	1	0	1	0	0	1	0
12	1	0	1	1	0	0	1
13	1	1	0	0	0	0	1
14	1	1	0	1	0	1	1
15	1	1	1	0	1	0	1
16	1	1	1	1	0	0	1

Fig. 2. Example data set

Step 8. The reduced data set consists of the representative instances in each partition, and is used for the instance-based learning. The proposed algorithm process is shown in Figure 1.

4 Example

The example data set consists of 16 instances, 6 attributes, and 1 class, and is presented in Figure 2.

Steps 1–4: Data partitioning

Step 1. Calculate the entropy of all attributes. To calculate the expected information of attribute A1, we use Equation (1) as follows:

For $A1 = 0$:

$$s_{11} = 6, s_{21} = 2, I(s_{11}, s_{21}) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} = 0.31 + 0.50 = 0.81$$

For $A1 = 1$:

$$s_{12} = 3, s_{22} = 5, I(s_{12}, s_{22}) = -\frac{3}{8} \log_2 \frac{3}{8} - \frac{5}{8} \log_2 \frac{5}{8} = 0.53 + 0.42 = 0.95$$

Using Equation (2), the entropy needed to classify a given sample if the samples are partitioned according to A1 is

$$E(A1) = \frac{8}{16} I(s_{11}, s_{21}) + \frac{8}{16} I(s_{12}, s_{22}) = \left(\frac{8}{16}\right) \times 0.81 + \left(\frac{8}{16}\right) \times 0.95 = 0.88.$$

Similarly, we can compute $E(A2) = 0.88$, $E(A3) = 0.88$, $E(A4) = 0.88$, $E(A5) = 0.82$, and $E(A6) = 0.60$.

Since A6 has the lowest entropy among the attributes, it is selected as the first attribute to use for partitioning. Next, divide the example data set by the values of attribute A6. The partitions are P1 and P2. After step 1, the result of partitioning is shown in Figure 3(a). The attribute values considered in the partition are marked in gray.

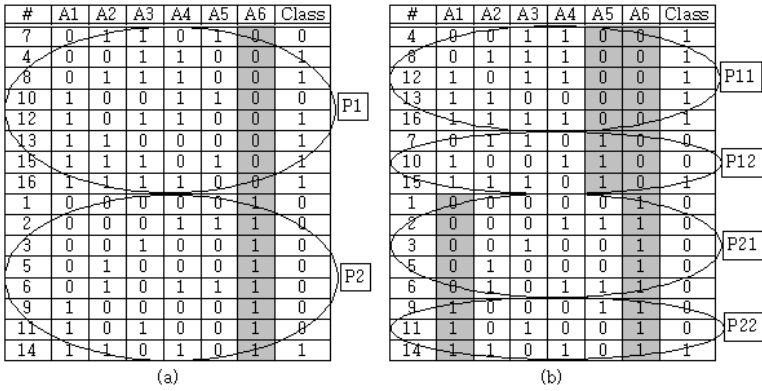


Fig. 3. (a) Partition sets after step1 (b) Partition sets after step 2

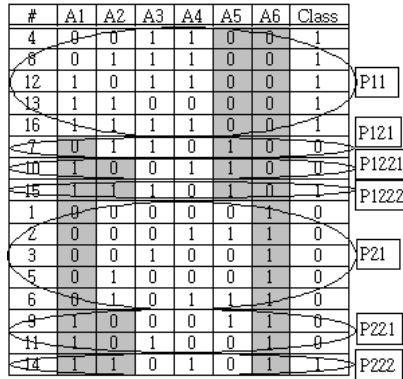


Fig. 4. Partition sets after data partitioning

Step 2. Partitions $P1$ and $P2$ are created. For each partition, calculate the entropy of the attributes that are not considered.

Because the entropy of attribute $A5$ is the smallest in partition $P1$, partition $P1$ is divided using attribute $A5$. Partition $P2$ is divided using attribute $A1$ for the same reason. The partitioning results after step 2 are shown in Figure 3(b). Partitions $P11$ and $P12$ are formed in partition $P1$, and partitions $P21$ and $P22$ are formed in partition $P2$.

Step 3. This partitioning process can continue until all partitions are pure or no further partitioning can be done.

Step 4. : Seven partition sets are composed in example: $\{P11, P121, P1221, P1222, P21, P221, \text{ and } P222\}$. The results after the data partitioning are shown in Figure 4.

Steps 5–8: Finding the representative instances

Step 5. Find the center instance of each partition using Equation (5).

When finding the center instance for partition P_{11} , only attributes A_1 and A_2 are used. Therefore, this method has the advantage of being able to locate the center instance faster, using only some and not all attributes. The center instance is decided based on the least Euclidean distance measure. For example, if we calculate the sum of the distances between instances for all instances in partition P_{11} , we get the following:

- The sum of the distances with instance #4 and the remaining instances = $\sqrt{1} + \sqrt{1} + \sqrt{2} + \sqrt{2} = 2 + 2\sqrt{2}$,
- The sum of the distances with instance #8 and the remaining instances = $\sqrt{1} + \sqrt{2} + \sqrt{1} + \sqrt{1} = 3 + \sqrt{2}$,
- The sum of the distances with instance #12 and the remaining instances = $\sqrt{1} + \sqrt{2} + \sqrt{1} + \sqrt{1} = 3 + \sqrt{2}$,
- The sum of the distances with instance #13 and the remaining instances = $\sqrt{2} + \sqrt{1} + \sqrt{1} + \sqrt{0} = 2 + \sqrt{2}$,
- The sum of the distances with instance #16 and the remaining instances = $\sqrt{2} + \sqrt{1} + \sqrt{1} + \sqrt{0} = 2 + \sqrt{2}$.

Therefore, the center instance in partition P_{11} becomes instance #13 or instance #16.

- Step 6.* In each partition, find the k instances nearest to the center instance. k is predefined at the data partitioning stage. For example, *one* instance is selected in partition P_{11} , and also *one* is selected in P_{21} . And no *one* instance is selected in the remaining partitions because they have only *one* or *two* instances.
- Step 7.* Find the representative instances in each partition. For example, *two* instances are selected in partition P_{11} : *one* center instance and *one* nearest instance. The result is the same in Partition P_{21} . Only *one* center instance is selected as the representative instance in the remaining partitions.
- Step 8.* The reduced data set consists of the representative instances in each partition. For example, the final reduced data set is shown in Figure 5.

#	A1	A2	A5	A6	Class
13	1	1	0	0	1
16	1	1	0	0	1
7	0	1	1	0	0
10	1	0	1	0	0
15	1	1	1	0	1
1	0	0	0	1	0
3	0	0	0	1	0
9	1	0	1	1	0
14	1	1	0	1	1

Fig. 5. Final reduced data set

5 Experimental Results

This data reduction algorithm was empirically compared with the k -nearest neighbor(k -nn) algorithm and Wai Lam’s PGF algorithm. Six data sets from

Table 1. Data sets

Data set	Number of instances	Number of attributes
Zoo	101	16
Audiology	226	63
Vote	435	17
Soybean	683	36
Credit	690	5
Mushroom	8124	22

the widely used UCI Database Repository were tested in the experiments [11]. The data sets used in the experiments are shown in Table 1. Experiments were performed on a PC with Pentium IV 3.0 GHz CPU and 512 MB of RAM. The implementation was done using Visual C++. For each data set, ten-fold cross-validation was used to estimate the average classification accuracy [10]. For each data set, we randomly partitioned the data into ten mutually exclusive subsets, S_1, S_2, \dots, S_{10} , each of approximately equal size. Training and testing were performed 10 times. The classifier of the first iteration was trained on subsets S_2, S_3, \dots, S_{10} and tested on S_1 , and the classifier of the second iteration was trained on subsets S_1, S_3, \dots, S_{10} and tested on S_2 , and so on.

The classification accuracy estimate is the overall number of correct classifications from the ten iterations, divided by the total number of instances in the data set. The data reduction rate is the product of the number of instances and the number of attributes in the reduced data set, divided by the product of the number of instances and the number of attributes in the original data set. Note that higher classification accuracy and a better data reduction rate imply better performance.

The detailed performance of each algorithm for each individual data set can be found in Table 2. Symbols “—” indicate that k -nn algorithm retains 100% of the original size (i.e., data reduction rate is 0%). The proposed algorithm removes 93.64% of the original size on average and the average accuracy is 89.42%.

In the Zoo data set, using the proposed algorithm, irrelevant 7 attributes among 16 attributes were removed, and the data set consisted of 14 representative instances. The data reduction rate of the Zoo data set is shown in Figure 6.

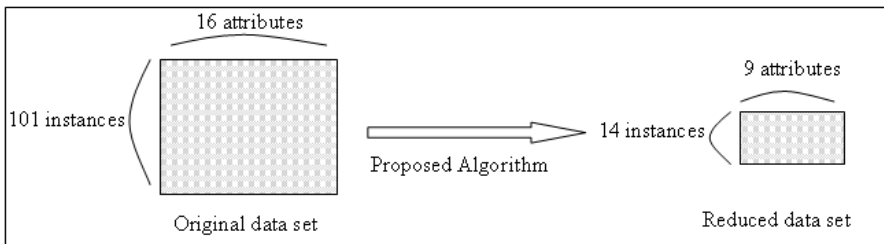
**Fig. 6.** Data reduction rate of the Zoo data set

Table 2. Classification accuracy(Accuracy) and data reduction rate(Size) for each data set

Data set		<i>k</i> -nn	PGF	Proposed Algorithm
Zoo	Accuracy	97.00	90.00	92.10
	Size	—	91.10	92.20
Audiology	Accuracy	76.10	67.20	78.50
	Size	—	87.00	90.10
Vote	Accuracy	95.50	92.60	92.80
	Size	—	93.90	94.00
Soybean	Accuracy	90.80	89.10	90.50
	Size	—	87.90	88.35
Credit	Accuracy	80.70	84.50	86.70
	Size	—	97.70	97.80
Mushroom	Accuracy	99.90	99.60	95.50
	Size	—	99.10	99.40
Average	Accuracy	89.67	87.17	89.42
	Size	—	92.78	93.64

Also, 2 – 7 attributes among 16 attributes were used to find the representative instances in each partition, and only 3.8 attributes were used on average.

In the Audiology, Vote, and Soybean data set, our proposed algorithm resulted in greater accuracy in classification and better reduction than the PGF algorithm. In particular, the proposed algorithm reduced 63 attributes to 30 in the Audiology data set.

In the Credit data set, the data reduction rate was high (97.8%). An initial 690 instances were reduced to 15 instances.

In the Mushroom data set, many attributes were regarded as irrelevant and removed. An initial 8124 instances were reduced to 178 instances, and 22 attributes were pared to 6 attributes.

When the proposed algorithm was compared with the *k*-nn algorithm, classification accuracy demonstrated similar results. In the case of the credit data set, the classification accuracy of the proposed algorithm was higher (86.7%). Also, when the proposed algorithm was compared with the PGF algorithm, the classification accuracy and data reduction rate were high in most cases. Through entropy-based partitioning that computing time is less than agglomerative hierarchical clustering method like PGF algorithm, data reduction was also achieved more quickly.

6 Conclusions

We have presented a new data reduction method for instance-based learning that integrates the strength of instance partitioning and attribute selection. Reducing the amount of data for instance-based learning reduces data storage

requirements, lowers computational costs, minimizes noise, and can facilitate a more rapid search.

Using the proposed algorithm, the initial data set is segmented into several partitions. Each partition is divided continuously based on entropy, and this partitioning process can continue until all partitions are pure or no further partitioning can be done. Finally, the homogeneity of instances in the same partition can be maintained. After dividing the initial data set, the attributes that are not used in the data partitioning stage are regarded as irrelevant and removed. Because irrelevant attributes can be removed, this method can find the representative instances of each partition more quickly than other methods.

Experimental results show that the proposed algorithm achieves a high data reduction rate as well as classification accuracy. The proposed algorithm can be employed to preprocess data used for data mining as well as in instance-based learning.

References

1. Liu, H., Hussain, F., Tan, C.L., Dash M.: Discretization: an enabling technique. *Data Mining Knowledge Discovery*. 6 (2002) 393-423
2. Cano, J.R., Herrera, F., Lozano M.: On the combination of evolutionary algorithms and stratified strategies for training set selection in data mining. *Applied Soft Computing*, In Press, Correted Proof, (2005)
3. Datta, P., Kibler, D.: Learning prototypical concept description. *Proceedings of the 12th International Conference on Machine Learning*. (1995) 158-166
4. Datta, P., Kibler, D.: Symbolic nearest mean classifier. *Proceedings of the 14th National Conference of Artificial Intelligence*. (1997) 82-87
5. Lam, W., Keung, C.L., Ling C.X.: Learning good prototypes for classification using filtering and abstraction of instances. *Pattern Recognition*, Vol. 35. (2002) 1491-1506
6. Sanchez, J.S.: High training set size reduction by space partitioning and prototype abstraction. *Pattern Recognition*, Vol. 37. (2004) 1561-1564
7. Dasarath, B.V.: *Nearest Neighbor Norms : NN Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, CA (1991)
8. Wilson, D.R., Martinez, T.R.: *Reduction Techniques for instance-based learning algorithms*. *Mach. Learning*. 38 (2000) 257-286
9. Cano, J.R, Herrera, F., Lozano, M.: Using evolutionary algorithms as instance selection for data reduction in kdd: an experimental study. *IEEE Transactions on Evolutionary Computation*. 7 (6) (2003) 561-575
10. Han, J., Kamber M.: *Data Mining : Concepts and Techniques*. Morgan Kaufman (2001)
11. Merz, C.J., Murphy, P.M. : *UCI Repository of Machine Learning Databases*, Internet: <http://www.ics.uci.edu/~mllearn/MLRepository.html>

Coordinated Inventory Models with Compensation Policy in a Three Level Supply Chain

Jeong Hun Lee and Il Kyeong Moon*

Department of Industrial Engineering, Pusan National University,
Busan, 609-735, Korea
Tel.: +82-51-510-2451; Fax: +82-51-512-7603
{jhlee, ikmoon}@pusan.ac.kr

Abstract. In this paper, we develop inventory models for the three level supply chain (one supplier, one warehouse, and one retailer) and consider the problem of determining the optimal integer multiple n of time interval, time interval between successive setups and orders in the coordinated inventory model. We consider three types of individual models (independent model, retailer's point of view model, and supplier's point of view model). The focus of this model is minimization of the coordinated total relevant cost, and then we apply the compensation policy for the benefits and losses to our coordinated inventory model. The optimal solution procedure for the developed model is derived and the effects of the compensation policy on the optimal results are studied with the help of numerical examples.

1 Introduction

While SCM is relatively new, the idea of coordinated model is not. The study of multi-echelon inventory/distribution systems began as early as 1960 by Clark and Scarf [5]. Since that time, many researchers have investigated multi-echelon inventory and distribution systems. Many researches have been aimed at coordinated model with two levels, while researchers who studied models with three levels are less. Erengüç *et al.* [7] point out that though a dominant firm in the supply chain usually tends to optimize locally with no regard to its impact on the other members of the chain, there are cases of such firms capable of fostering more cooperative agreements in the chain. An empirical study on buyer-supplier relationship highlighted the importance of strong linkages for efficient JIT operations [3]. They called for replacing the traditional adversarial roles between buyers and sellers with mutual cooperation. Kang *et al.* [13] have reviewed past and present supply chain models and then analyzed those in view of environment factors, operations, solution approaches. Goyal [10] presented an integrated inventory model for a single supplier-single customer problem. Banerjee [1] presented a joint economic-lot-size model where a vendor produces to order for a purchaser on a lot-for-lot basis under deterministic conditions. Goyal [11] further generalized Banerjee [1]'s model by relaxing the assumption of the

* Corresponding author.

lot-for-lot policy of the vendor. As a result of using the approach suggested by Goyal [11], significant reduction in inventory cost can be achieved. Several researchers have shown that one partner's gain may exceed the other partners' loss in integrated models. Thus, the net benefit can be shared by both parties in some equitable fashion [12]. Eum *et al.* [8] proposed a new allocation policy considering buyers' demands using the neural network theory. Douglas and Paul [6] defined three categories of operational coordination (buyer-vendor coordination, production-distribution coordination and inventory-distribution coordination).

The value of information sharing among supply chain players has received much attention from researchers [4, 14]. They showed that using the information on the outstanding orders of the products resulted in improvement in system performance in a two-product model. Bourland *et al.* [2] demonstrated the value of obtaining demand information at the retailers. Gavirneni *et al.* [9] captured the value of information flow in a two-echelon capacitated model. Recently, Lee *et al.* [14] addressed the issue of quantifying the benefits of sharing information and identifying the drivers of the magnitude of these drivers.

Most of the works in the literature consider two level supply chain. But our work considers three level supply chain (one supplier, one warehouse and one retailer). While they provided the joint analysis of two level supply chain (the supplier and the buyer), our paper aims to study coordinated analysis among the supplier, the warehouse and the retailer, and strategies to encourage coordination among supply chain partners.

The rest of this paper is organized as follows. We introduce the problem in Section 2. In Section 3 we optimize for individual models. Section 4 establishes the coordinated model and develops the procedure for the minimization of the total cost in a three level supply chain. Section 5 develops compensation policies for benefits and losses using the developed models. We present numerical examples in Section 6 and conclude in Section 7.

2 Problem Definition

We consider a supply chain with three levels. Suppose that a retailer periodically orders some quantity (Q) of an inventory item from a warehouse, while a warehouse periodically orders integer multiple of the retailer's order quantity ($n \cdot Q$) of item from a supplier. Upon receipt of an order, the supplier produces the integer multiple quantity of the item. But the warehouse ships some quantity (Q) to the retailer during the multiple times (n). In addition to the deterministic conditions, we assume that there are no other warehouses and retailers for this item and the supplier is the sole manufacturer. Fig. 1 shows the inventory time plots for $n=3$. The retailer's time interval between successive orders is T_R , and the supplier and the warehouse's time interval between successive setups and orders are $T_S=3T_R$ and $T_W=3T_R$, respectively. At the end of time interval in the supplier, it delivers the completed lot to the warehouse. At the beginning of time interval in the warehouse, it directly delivers as the retailer's order quantity (Q) to the retailer (similar to cross-docking, a process in which product is exchanged between trucks so that each truck going to a retailer store has products from different suppliers). In the remaining time interval, it delivers two more times with the same quantity to the retailer.

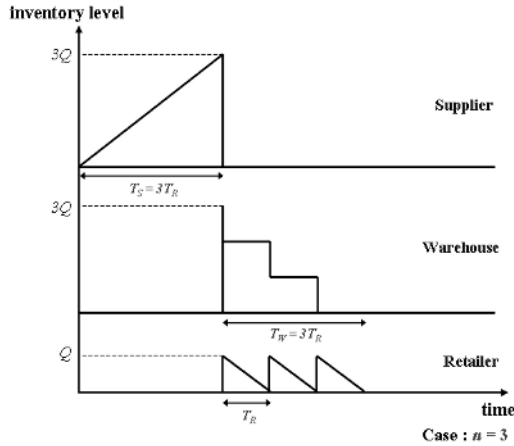


Fig. 1. Graphical representation of the model ($n = 3$)

The following assumptions are made to develop the models:

- (1) The demand is deterministic and constant.
- (2) Supplier, warehouse, retailer's lead-time are either zero or replenishment is instantaneous.
- (3) The holding cost values are $h_R > h_W$. Because the warehouse takes charge of the storage and distribution professionally, the warehouse's holding cost is less than the retailer.
- (4) The supplier's time interval between setups and the warehouse's time interval between orders are integer ($n > 1$) multiple of the retailer's time interval between orders.
- (5) Shortages and backlogs are not allowed.
- (6) Supplier and retailer's inventory policies can be described by simple EOQ inventory model.

The following notations are used in developing the models:

- D : annual demand for the item
- S : supplier's setup cost per setup
- A_W : warehouse's ordering cost per order
- A_R : retailer's ordering cost per order
- T_S : time interval between successive setups at supplier
- T_W : time interval between successive orders at warehouse
- T_R : time interval between successive orders at retailer
- h_S : supplier's holding cost per unit per unit time
- h_W : warehouse's holding cost per unit per unit time
- h_R : retailer's holding cost per unit per unit time
- n : positive integer number ($n > 1$)

3 Individual Model

We consider three types of individual models. Firstly, we formulate an independent individual model. In the independent individual model, manufacturing and ordering policies are independent. Secondly, we develop an individual model from retailer’s point of view. In this model, the retailer is a decision-maker. Therefore, the other parties follow the retailer's ordering policy. For example, department stores decide the ordering policies regardless of the other parties, because they have a power in the marketplace. Thirdly, we consider an individual model from supplier’s point of view. This model is opposite to the retailer's point of view model. Because supplier is a decision-maker, the warehouse and retailer decide ordering policies according to the supplier's decision. For instance, the high-technology products are made by the supplier's decision regardless of the warehouse and retailer's order. Because the supplier and retailer’s inventory policies can be described by simple EOQ, we can easily derive the optimal policies. The results of individual optimization are summarized in Table 1.

Table 1. Summary of total costs and individual optimal policies

	Supplier	Warehouse	Retailer
Independent	$TC_s(T_s) = \frac{S}{T_s} + \frac{DT_s}{2} h_s$ $T_s^* = \sqrt{\frac{2S}{Dh_s}}$ $TC_s(T_s^*) = \sqrt{2DS}h_s$	$TC_w(n, T_w) = \frac{A_w}{T_w} + \frac{(n-1)DT_w}{2} h_w$ $n^* = 1, T_w^* = T_s^*$ $TC_w(n^*, T_w^*) = \frac{A_w}{T_s^*}$	$TC_r(T_r) = \frac{A_r}{T_r} + \frac{DT_r}{2} h_r$ $T_r^* = \sqrt{\frac{2A_r}{Dh_r}}$ $TC_r(T_r^*) = \sqrt{2DA_r}h_r$
Retailer’s point of view	$TC_s(T_s) = \frac{S}{T_s} + \frac{DT_s}{2} h_s$ $T_s^* = T_r^*$ $TC_s(T_s^*) = \frac{S}{T_r^*} + \frac{DT_r^*}{2} h$	$TC_w(n, T_w) = \frac{A_w}{T_w} + \frac{(n-1)DT_w}{2} h_w$ $n^* = 1, T_w^* = T_r^*$ $TC_w(n^*, T_w^*) = \frac{A_w}{T_r^*}$	$TC_r(T_r) = \frac{A_r}{T_r} + \frac{DT_r}{2} h_r$ $T_r^* = \sqrt{\frac{2A_r}{Dh_r}}$ $TC_r(T_r^*) = \sqrt{2DA_r}h_r$
Supplier’s point of view	$TC_s(T_s) = \frac{S}{T_s} + \frac{DT_s}{2} h_s$ $T_s^* = \sqrt{\frac{2S}{Dh_s}}$ $TC_s(T_s^*) = \sqrt{2DS}h_s$	$TC_w(n, T_w) = \frac{A_w}{T_w} + \frac{(n-1)DT_w}{2} h_w$ $n^* = 1, T_w^* = T_s^*$ $TC_w(n^*, T_w^*) = \frac{A_w}{T_s^*}$	$TC_r(T_r) = \frac{A_r}{T_r} + \frac{DT_r}{2} h_r$ $T_r^* = T_s^*$ $TC_r(T_r^*) = \frac{A_r}{T_s^*} + \frac{DT_s^*}{2} h$

The stock in the warehouse is depleted according to the demand and supply. If the warehouse is replenished at a time interval of T_w and the quantity received can satisfy multiple orders, then the total cost per unit time is given by

$$TC_w(n, T_w) = \frac{A_w}{T_w} + \frac{(n-1)DT_w}{2} h_w \tag{3.1}$$

The objective is to find the optimal values of n and T_w which minimize $TC_w(n, T_w)$. Since n is a positive integer and T_w is a real number, we can optimize the total cost per unit time as given below:

For any given $n (\geq 1)$, as the second order derivative of $TC_w(n, T_w)$ is always positive, the necessary condition for the minimum of $TC_w(n, T_w)$ is given by

$$\frac{\partial TC_w}{\partial T_w} = \frac{(n-1)Dh_w}{2} - \frac{A_w}{T_w^2} = 0 \tag{3.2}$$

Solving equation (3.2) we get

$$T_w^* = \sqrt{\frac{2A_w}{(n-1)Dh_w}} \tag{3.3}$$

Substituting T_w from equation (3.3) into equation (3.1), the total cost per unit time can be found for any given n . It is to be observed that there exists a unique optimal solution (n^*, T_w^*) as $TC_w(n, T_w)$ is convex for any given n .

$$TC_w(n) = \sqrt{2(n-1)DA_w h_w} \tag{3.4}$$

Minimizing $TC_w(n)$ is equivalent to minimizing

$$(TC_w(n))^2 = 2(n-1)DA_w h_w \tag{3.5}$$

We define $Y(n)$ which enables our problem to be equivalent to the minimization of $Y(n)$.

$$Y(n) = 2(n-1)DA_w h_w \tag{3.6}$$

However, $Y(n)$ is a linear increasing function which depends on n . Therefore, the optimal minimum value of n is always 1. It means that the supplier directly delivers order quantity to the retailer. The role of the warehouse is similar to the cross-docking (CD) system. Hence, the warehouse is spending only the ordering cost, and the optimal value of T_w is equal to T_s^* and T_R^* .

4 Coordinated Model

The relevant total cost of the coordinated model for the supplier, the warehouse and the retailer can be derived by adding the individual total costs per unit time from the previous section.

$$CTC(n, T_R) = \frac{A_R}{T_R} + \frac{A_w}{nT_R} + \frac{S}{nT_R} + \frac{DT_R h_R}{2} + \frac{(n-1)DT_R h_w}{2} + \frac{nDT_R h_S}{2} \tag{4.1}$$

where $T_w = n \cdot T_R$ and $T_S = n \cdot T_R$

The optimal values of n and T_R can be obtained using the following propositions.

Proposition 1: For any given $n (\geq 1)$, the time interval between successive setups and reorders in the coordinated model can be determined uniquely.

Proof: Differentiating equation (4.1) with respect to T_R , we get

$$\frac{\partial CTC}{\partial T_R} = \frac{D(h_r + nh_w - h_w + nh_s)}{2} - \frac{nA_r + A_w + S}{nT_R^2} \tag{4.2}$$

Differentiating equation (4.2) again with respect to T_R , we get

$$\frac{\partial^2 CTC}{\partial T_R^2} = \frac{2(nA_r + A_w + S)}{nT_R^3} > 0, \forall n \geq 1 \tag{4.3}$$

Hence $CTC(n, T_R)$ is convex in T_R when n is given. Therefore, there exists a unique solution of the equation $\partial CTC(n, T_R) / \partial T_R = 0$ which yields

$$T_R^* = \sqrt{\frac{2(nA_r + A_w + S)}{nD\{h_r + (n-1)h_w + nh_s\}}} \tag{4.4}$$

Substituting T_R^* into equation (4.1), we obtain the minimum total cost of the coordinated model as follows:

$$CTC(n) = \sqrt{\frac{2D\{h_r - h_w + n(h_w + h_s)\}(nA_r + A_w + S)}{n}} \tag{4.5}$$

We can find the optimal value of n using the following proposition.

Proposition 2: The optimal value of n satisfies the following inequality.

$$n^*(n^* - 1) \leq \frac{(h_r - h_w)(S + A_w)}{A_r(h_s + h_w)} \leq n^*(n^* + 1)$$

Proof: Minimizing $CTC(n)$ is equivalent to minimizing

$$(CTC(n))^2 = 2D\{(h_r - h_w)(A_r + \frac{S + A_w}{n}) + (h_s + h_w)(nA_r + A_w + S)\} \tag{4.6}$$

After ignoring the terms on the right hand side of equation (4.6) which are independent of n , we define $Z(n)$ as follows:

$$Z(n) = \frac{(h_r - h_w)(S + A_w)}{n} + nA_r(h_s + h_w) \tag{4.7}$$

The optimal value of $n = n^*$ is obtained when

$$Z(n^*) \leq Z(n^* - 1) \text{ and } Z(n^*) \leq Z(n^* + 1) \tag{4.8}$$

We get the following inequalities from (4.8)

$$\begin{aligned} \frac{(h_r - h_w)(S + A_w)}{n^*} + n^*A_r(h_s + h_w) &\leq \frac{(h_r - h_w)(S + A_w)}{n^* - 1} + (n^* - 1)A_r(h_s + h_w) \text{ and} \\ \frac{(h_r - h_w)(S + A_w)}{n^*} + n^*A_r(h_s + h_w) &\leq \frac{(h_r - h_w)(S + A_w)}{n^* + 1} + (n^* + 1)A_r(h_s + h_w) \end{aligned} \tag{4.9}$$

Accordingly, it follows that

$$\begin{aligned}
 A_R(h_s + h_w) &\leq \frac{1}{n^*(n^* - 1)}(h_R - h_w)(S + A_w) \text{ and} \\
 A_R(h_s + h_w) &\geq \frac{1}{n^*(n^* + 1)}(h_R - h_w)(S + A_w)
 \end{aligned}
 \tag{4.10}$$

The following condition is obtained from equation (4.10):

$$n^*(n^* - 1) \leq \frac{(h_R - h_w)(S + A_w)}{A_R(h_s + h_w)} \leq n^*(n^* + 1)
 \tag{4.11}$$

< Procedure for finding n^* and T_R^* >

Step 1: Determine

$$\frac{(h_R - h_w)(S + A_w)}{A_R(h_s + h_w)}$$

If n is greater than 1 in inequality (4.11), set $n^* = n$. Otherwise, set $n^* = 1$. Go to Step 2.

Step 2: Determine the optimal value of T_R using equation (4.4).

5 Compensation Policy

Several researchers have shown that one partner’s gain may exceed the other partner’s loss in the integrated model [9, 14]. Thus, the net benefit should be shared among parties (the supplier, the warehouse and the retailer) in some equitable fashion. We propose a compensation policy that shares benefits and losses according to the ratio of individual models’ total cost per unit time. This method extends Goyal [10]’s method to the three level supply chain.

Applying Goyal’s method to our coordinated model, we get

$$\begin{aligned}
 Z_s &= \frac{TC_s(T_s^*)}{TC_s(T_s^*) + TC_w(n^*, T_w^*) + TC_R(T_R^*)} \\
 \text{Cost of supplier} &= Z_s \cdot CTC(n^*, T_R^*)
 \end{aligned}
 \tag{5.1}$$

$$\begin{aligned}
 Z_w &= \frac{TC_w(n^*, T_w^*)}{TC_s(T_s^*) + TC_w(n^*, T_w^*) + TC_R(T_R^*)} \\
 \text{Cost of warehouse} &= Z_w \cdot CTC(n^*, T_R^*)
 \end{aligned}
 \tag{5.2}$$

$$\begin{aligned}
 Z_R &= \frac{TC_R(T_R^*)}{TC_s(T_s^*) + TC_w(n^*, T_w^*) + TC_R(T_R^*)} \\
 \text{Cost of retailer} &= Z_R \cdot CTC(n^*, T_R^*)
 \end{aligned}
 \tag{5.3}$$

Note that $Z_s + Z_w + Z_R = 1$

6 Numerical Examples

For numerical examples, we use the following data:

$$D = 10,000 \text{ unit/year}, S = \$400/\text{setup}, A_W = \$200/\text{order}, A_R = \$50/\text{order}$$

$$h_S = \$3/\text{unit/year}, h_W = \$3/\text{unit/year}, h_R = \$3/\text{unit/year}$$

The optimal values of n , T_R and total cost for the individual models and the coordinated model are summarized in Table 2.

Table 2. Summary of results

	Individual models			Coordinated model
	Independent	Retailer's point of view	Supplier's point of view	
Supplier's setup interval	0.1633 year	0.0447 year	0.1633 year	0.1581
Warehouse's order interval	0.1633 (n=1)	0.0447 (n=1)	0.1633 (n=1)	0.1581 (n=3)
Retailer's order interval	0.0447	0.0447	0.1633	0.0527
Supplier's annual cost	\$4,898.98	\$9,615.34	\$4,898.98	\$4,901.53
Warehouse's annual cost	\$1,224.75	\$4,472.17	\$1,224.75	\$2,319.00
Retailer's annual cost	\$2,236.07	\$2,236.07	\$4,388.68	\$2,266.30
Total cost	\$8,359.80	\$16,323.58	\$10,512.41	\$9,486.83

Fig. 2 shows that the total cost function $CTC(n, T_R)$ is a convex function in n and T_R and a typical configuration of the surface.

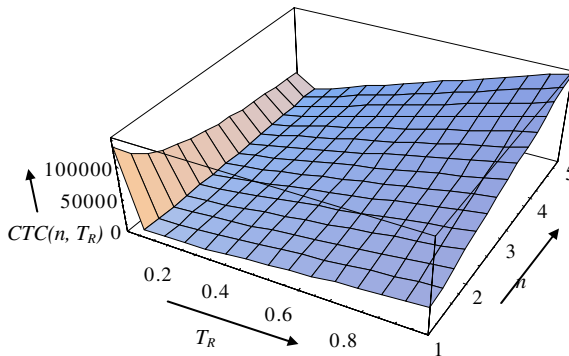


Fig. 2. Graphical Representation of $CTC(n, T_R)$

If we coordinate the three level supply chain, we can reduce \$6,836.75 (approximately 42%) of the total cost against retailer's point of view model. Therefore, we need to share the benefits. Applying the compensation policy using equations (5.1), (5.2), and (5.3), we get the following results:

$$Z_s = 0.5890, \text{ Cost of supplier} = \$5,587.74$$

$$Z_w = 0.2740, \text{ Cost of warehouse} = \$2,599.39$$

$$Z_r = 0.1370, \text{ Cost of retailer} = \$1,299.70$$

Fig. 3 summarizes the process of applying the compensation policy and information sharing.

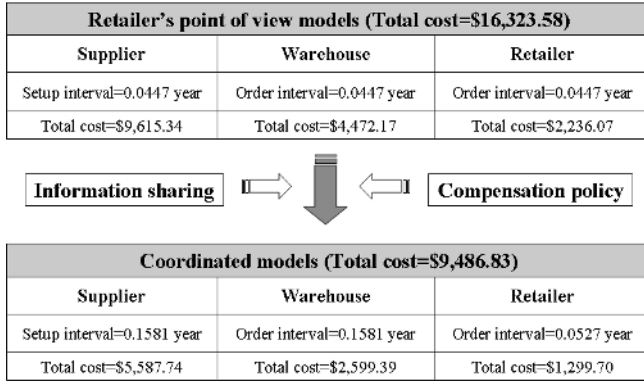


Fig. 3. Graphical illustration of the solutions

7 Conclusions

We developed an inventory model for a three level supply chain (one supplier, one warehouse, and one retailer). We proposed a procedure for determining the optimal value of n and T_R for the coordinated model. The compensation policy gives better results than individual models in terms of the total cost per unit time. The total cost per unit time obtained by the coordinated model with compensation policy has been reduced significantly compared to the individual models. We may develop other types of compensation policy (i.e. price quantity discounts policy). In addition, our model can be extended to the case with multiple suppliers, one warehouse, and multiple retailers. Finally, it must be an interesting extension if one could develop the model by relaxing the assumptions of deterministic demand and lead time.

Acknowledgements

This work was supported by the Regional Research Centers Program(Research Center for Logistics Information Technology), granted by the Korean Ministry of Education & Human Resources Development.

References

1. Banerjee, A., A joint economic-lot-size model for purchaser and vendor, *Decision Sciences*, 17 (1986) 292-311
2. Bourland, K., Powell, S. and Pyke, D., Exploring timely demand information to reduce inventories, *European Journal of Operational Research*, 92 (1996) 239-253

3. Chapman, S. N. and Carter, P. L., Supplier/customer inventory relationships under just in time, *Decision Science*, 21 (1990) 35-51
4. Chopra, S. and Meindl, P., *Supply Chain Management: Strategy, Planning and Operation*, Prentice-Hall, Inc (2001)
5. Clark, A. J. and Scarf, H., Optimal policies for a multi-echelon inventory problem, *Management Science*, 6 (1960) 475-490
6. Douglas J. T. and Paul, M. G., Coordinated supply chain management, *European Journal of Operational Research*, 94 (1996) 1-15
7. Erengüç, S. S., Simpson, N. C. and Vakharia, A. J., Integrated production/distribution planning in supply chains: an invited review, *European Journal of Operational Research*, 115 (1999) 219-236
8. Eum, S. C., Lee Y. H. and Jung J. W., A producer's allocation policy considering buyers' demands in the supply chain, *Journal of the Korean Institute of Industrial Engineers*, 31 (2005) 210-218
9. Gavirneni, S., Kapuscinski, R. and Tayur, S., Value of information in capacitated supply chains, *Management Science*, 45 (1999) 16-24
10. Goyal, S. K., An integrated inventory model for a single supplier-single customer problem, *International Journal of Production Research*, 15 (1976) 107-111
11. Goyal, S. K., A joint economic-lot-size model for purchaser and vendor: a comment, *Decision Sciences*, 19 (1988) 236-241
12. Goyal, S. K. and Gupta, Y. P., Integrated inventory models: the buyer-vendor coordination, *European Journal of Operational Research*, 41 (1989) 261-269
13. Kang, K. H., Lee B. K. and Lee Y. H., A review of current status and analysis in supply chain moderling, *Journal of the Korean Institute of Industrial Engineers*, 30 (2004) 224-240
14. Lee, H. L., So, K. C. and Tang, C. S., The value of information sharing in a two-level supply chain, *Management Science*, 46 (2000) 626-643

Using Constraint Satisfaction Approach to Solve the Capacity Allocation Problem for Photolithography Area

Shu-Hsing Chung¹, Chun-Ying Huang¹ and Amy Hsin-I Lee²

¹ Department of Industrial Engineering and Management, National Chiao Tung University,
No. 1001, Ta Hsueh Road, Hsinchu, Taiwan, R.O.C.
shchung@mail.nctu.edu.tw, cyhuang.iem90g@nctu.edu.tw

² Department of Industrial Management, Chung Hua University,
No. 707, Sec.2, Wu Fu Road, Hsinchu, Taiwan, R.O.C.
amylee@chu.edu.tw

Abstract. This paper addresses the capacity allocation problem for photolithography area (CAPPA) under an advanced technology environment. The CAPPA problem has two characteristics: process window and machine dedication. Process window means that a wafer needs to be processed on machines that can satisfy its process capability (process specification). Machine dedication means that after the first critical layer of a wafer lot is being processed on a certain machine, subsequent critical layers of this lot must be processed on the same machine to ensure good quality of final products. A production plan, constructed without considering the above two characteristics, is difficult to execute and to achieve its production targets. Thus, we model the CAPPA problem as a constraint satisfaction problem (CSP), which uses an efficient search algorithm to obtain a feasible solution. Additionally, we propose an upper bound of load unbalance estimation to reduce the search space of CSP for searching an optimal solution. Experimental results show that the proposed model is useful in solving the CAPPA problem in an efficient way.

1 Introduction

Due to its diverse characteristics, such as reentry process, time-constrained operation and different batch sizes for machines, wafer fabrication has received a lot of research attention, especially in photolithography process [13, 14, 7, 10]. The photolithography process uses masks to transfer circuit patterns onto a wafer, and the etching process forms tangible circuit patterns onto the wafer chip. With the required number of processes in the photolithography, integrated circuitry products with preset functions are developed on the wafer.

As wafer fabrication technology advances from micrometer level to nanometer level, more stringent machine selection restrictions, the so-called process window control and machine dedication control, are imposed on the production management of photolithography area for wafer lots.

Process window constraint, also called equipment constraint, is related to the strict limitation to the choice of a machine to process higher-end fabrication technology in the process of a wafer lot to meet increasingly narrower line width, distance between lines, and tolerance limit. In other words, wafer lots could only be processed on

machines that meet certain process capability (process recipe or process specification). On the contrary, wafers that only need a lower-end fabrication technology have less stringent machine selection restriction. Due to the difference in adjustable ability among photolithography machines regarding process recipes, functions of various machines in fact vary to a certain extent even though they are grouped in the same workstation. Hence, the situation is that some machines can handle more process capabilities (simultaneously handle higher- and lower-end fabrication technology) while other machines can handle less process capabilities (only handle lower-end fabrication technology). Some related studies are as follows. Leachman and Carmon [9] and Hung and Cheng [5] use linear programming to obtain a production plan for maximizing the profit. Toktay and Uzsoy [12] transform the capacity allocation problem with machines' capabilities constraint into maximum flow problem. However, only a single product type is considered in the study.

Machine dedication constraint considers layer-by-layer process on wafers, in which the circuit patterns in the layers can be correctly connected in order to provide particular functions. If electrical circuits among the layers cannot be aligned and connected, this will cause defective products. The alignment precision provided by different machines varies to a certain extent, even for machines of the same model, due to some differences, which are referred to as machine difference. It has been stipulated that when the first critical operation of a wafer lot is done on a particular machine, the rest of its subsequent critical processes will need to be processed by the very same machine to avoid the increase in defective rate due to machine difference. A related study was done by Akçali *et al.* [1], in which a study was conducted on the correlation between photolithography process characteristics and production cycle time using a simulation model, and machine dedication policy was set as one of the experiment factors. Experimental results indicate that dedicated assignment policy has a remarkable impact on cycle time.

With advanced fabrication technology, the impact of process window and machine dedication constraints on wafer fabrication is increasingly evident. Capacity requirement planning is difficult because wafer fabrication has special characteristics of reentry and long cycle time, and the number of layers of products, required process window, number and distribution of critical layers are different. As a result, the effectiveness of production planning and scheduling system is seriously impacted if the constraints of process window and machine dedication are not considered.

Up to now, the CAPPA problem has not been tackled except by Chung *et al.* [3], in which a mixed integer-linear programming (MILP) model is devised. However, a practical scale-sized problem may take an exponential time. In this research, we adopt the efficient constraint satisfaction approach, which treats load unbalance among machines as one of the constraints and obtain the optimal solution by constantly narrowing down the upper bound of the load unbalance. Because a relatively large amount of settings and long solving process may still be required, a load unbalance estimation to reduce the search space is also applied.

In section 2, the MILP model, the constraint satisfaction problem, and the load unbalance estimation are introduced. Section 3 demonstrates the effectiveness of proposed model. Section 4 uses a real-world case to show the applicability of proposed model. In the last section, the research results are summarized.

2 Model Development

Indices:

- i Index of order number, where $i = 1, \dots, I$.
- j Index of layer number, where $j = 1, \dots, J_i$.
- k Index of machine number in photolithography area, where $k = 1, \dots, K$.
- l Index of process capability number, where $l = 1, \dots, L$.
- t Index of planning period, where $t = 1, \dots, T$.
- r Index of ranking, where $r = 1, \dots, L$.

Parameters:

- A_{kl} = 1 if machine k has process capability l ; 0, otherwise.
- AC_{kt} Available capacity of machine k in planning period t .
- AL_t Average capacity loading of workstation in planning period t .
- CL_{ij} = 1 if layer j of order i is a critical layer; 0, otherwise.
- CR_{ijl} = 1 if layer j of order i has a load on process capability l ; 0, otherwise.
- CS_{rt} Cumulative available capacity from the first to the r -th rank process capability.
- DC_{lt} Capacity requirement of process capability l in planning period t .
- DS_{lt} Ratio of capacity requirement to available capacity of process capability l in planning period t .
- J_i Number of photolithography operations for order i .
- LT_{ijt} = 1 if layer j of order i has a load in planning period t ; 0, otherwise.
- ML_t The maximum loading level among machines in planning period t .
- p_{ij} Processing time of layer j of order i .
- $SQ(r)$ Function of the processing capacity of the r -th rank.
- SC_{lt} Available capacity of process capability l in planning period t .

Decision Variables:

- b_{ik} = 1 if the first critical layer of order i is assigned to machine k ; 0, otherwise.
- u_{kt}^+ Positive difference between utilization rate of machine k and average utilization rate of the entire workstation that machine k belongs to in planning period t .
- u_{kt}^- Negative difference between utilization rate of machine k and average utilization rate of the entire workstation that machine k belongs to in planning period t .
- x_{ijk} = 1 if layer j of order i is assigned to machine k ; 0, otherwise.

2.1 Mixed Integer-Linear Programming Model

To solve the CAPP problem, we need to know the load occurrence time of photolithography workstation for each layer of each order in the production system. An interview with several semiconductor fabricators found out that X-factor, the ratio of remaining time before delivery to processing time of an order, is used as a reference for controlling the production progress to make sure that the delivery of orders can be accomplished on time (see also [8]). With the information of X-factor, processing time

of an order, production plan and WIP level, the loading occurrence time of each order (LT_{ijt}) can be estimated. A MILP model is constructed as follows:

$$\text{Minimize } \sum_{t=1}^T \sum_{k=1}^K (u_{kt}^+ + u_{kt}^-) \tag{1}$$

Subject to

$$\sum_i \sum_k \sum_l \sum_j (x_{ijk} A_{kl} CR_{ijl} LT_{ijt}) = \sum_i \sum_l \sum_j (CR_{ijl} LT_{ijt}) \quad , \text{ for all } i \tag{2}$$

$$\sum_k x_{ijk} = 1 \quad , \text{ for all } i, j \tag{3}$$

$$\sum_i \sum_l \sum_j (x_{ijk} CL_{ij} CR_{ijl} LT_{ijt}) = b_{ik} \times \sum_i \sum_l \sum_j (CL_{ij} CR_{ijl} LT_{ijt}) \quad , \text{ for all } i, k \tag{4}$$

$$u_{kt}^+ - u_{kt}^- = \sum_i \sum_j \sum_l (x_{ijk} P_{ij} A_{kl} CR_{ijl} LT_{ijt}) / AC_{kt} - AL_t \quad , \text{ for all } t, k \tag{5}$$

$$u_{kt}^+ \leq 1 - AL_t \quad , \text{ for all } t, k \tag{6}$$

$$u_{kt}^- \leq AL_t \quad , \text{ for all } t, k \tag{7}$$

$$x_{ijk} \in \{0,1\} \quad , \text{ for all } i, j, k \tag{8}$$

$$b_{ik} \in \{0,1\} \quad , \text{ for all } i, k \tag{9}$$

$$u_{kt}^+ \geq 0 \quad , \text{ for all } t, k \tag{10}$$

$$u_{kt}^- \geq 0 \quad , \text{ for all } t, k \tag{11}$$

The objective function (1) is to balance the capacity utilization rates among machines. Constraint (2) ensures that each layer of an order, including new release orders and WIP orders, must be assigned to a machine k if it has a capacity request in this planning horizon. In the machine assignment, process window constraint must be considered. Constraint (3) is to make sure that each layer of an order can only be assigned to a particular single machine. Constraint (4) states the machine dedication control. If the first critical layer of order i is assigned to machine k for process, b_{ik} is set to one. Note that the orders in a planning horizon can either be orders planned to release or WIP orders that were released to shop floor in the previous planning horizon. Therefore, b_{ik} is a decision variable if the order is a planned-to-release order or a WIP order which its first critical layer has not been decided on a particular machine in previous planning horizon, and is a known parameter if the order is a WIP order which its first critical layer has been decided to process on machine k, but unfinished, in the previous planning horizon. Constraint (5) calculates the difference between the utilization rate of machine k and the average utilization rate of the entire workstation in each period of the planning horizon. The detail definition of AL_t is shown as equation (14) and (15) in section 2.3. Constraint (6) and (7) limit the upper value of u_{kt}^+ and u_{kt}^- , respectively.

2.2 Constraint Satisfaction Problem (CSP)

Constraint satisfaction problem (CSP) searches for a feasible solution which satisfies all constraints under a finite domain of variables. CSP originated from artificial intelligence (AI) in computer science. Through consistency checking techniques, constraint propagation and intelligent search algorithms, CSP has a relatively high solving efficiency in a combinatorial optimization problem, and it has been widely

applied in many research fields, such as vehicle routing related problem, production scheduling, facility layout and resource allocation [2, 11, 4].

Although CSP algorithm primarily aims to derive a feasible solution, it can be adjusted to search for an optimal solution. A feasible solution is generated by CSP first, then the feasible solution is set as the upper bound of the objective function (for a minimization problem), and such a relationship is treated as a constraint to solve the new CSP. With the continuation of lowering the upper bound of the objective function, the optimal solution can finally be obtained as the solution of the previous CSP when the current CSP can no longer be solved [2]. In other words, by deleting the objective function and solving the constraints part, we could convert the problem into a CSP. If the objective function is added into the constraints by setting its upper bound, an optimal solution for the CAPP problem can be obtained by constantly reducing the upper bound of the objective function (E), as shown by equation (12).

$$\sum_t \sum_k (u_{kt}^+ + u_{kt}^-) \leq E \tag{12}$$

Chung *et al.* [3] stated that loading balance is a critical factor for maintaining stability of production cycle time. Thus, we believe that the balance of loading among machines is more suitable than the emphasis of the minimization of the sum of the differences among machine utilization rates in a workstation. As a result, constraint (13) replaces constraint (12) in the solving of the CAPP by CSP. For a more convenient explanation, we refer a CSP to search for a feasible solution as I-CSP model, and a CSP to search for an optimal solution as O-CSP model.

$$u_{kt}^+ + u_{kt}^- \leq E_t, \text{ for each } k, \text{ each } t \tag{13}$$

When CSP is applied to generate an optimal solution for the CAPP problem (i.e. O-CSP model), the number of iterations for upper bound of load unbalance is difficult to estimate. In consequence, the expectation of a fast-solving and efficient algorithm from CSP may not be attained. Hence, we present a heuristic method to estimate the value of E_t (upper bound of load unbalance, UBLU). With a good estimation of the upper bound of load unbalance, the search space of O-CSP can be reduced, and the efficiency and quality of solution from CSP to solve the CAPP problem can be increased.

2.3 Upper Bound of Load Unbalance (UBLU) Estimation

Conventionally, the average load level (AL_t) of a workstation in a planning period is obtained by dividing total load by the number of machines, and the maximum loading level (ML_t) among the machines is assumed equal to the average loading level (AL_t). This calculation is based on the assumption that all machines are identical, that is, machines have identical process capability.

Since the types and amount of process capabilities are not exactly the same in the CAPP problem, the maximum loading level may not equal to the average loading level. Therefore, we propose a two-phase capacity demand-supply assessment, which includes an independent and a dependent assessment, to estimate the maximum loading level among the machines. The results are utilized as the basis for setting the UBLU in equation (13). The concept is described as follows:

Phase I. Independent capacity supply assessment

The independent assessment examines whether capacity requirement is less than capacity supply for each process capability l . Capacity supply is the sum of the maximum loading level (ML_l) of each machine to handle this process capability, and the initial value of ML_l is set to be the average capacity loading (AL_l) in a workstation. If capacity demand is less than supply, the independent assessment is passed, and we can go to the dependent assessment. Otherwise, the maximum loading level of all machines needs to be raised to satisfy the capacity requirement of process capability l .

Phase II. Dependent capacity supply assessment

Since Phase I evaluates the capacity demand-supply without considering the fact that machines may possess several process capabilities, this phase uses an iterative calculation to assess whether the overall capacity supply is sufficient based on the maximum loading level obtained from Phase I. First, the ratios (DS_{lt}) of capacity requirement to capacity supply of each process capability are ranked from large to small, and the sequence is $SQ(r)$. Then, whether cumulative capacity requirement is less than cumulative capacity supply is examined according to the ranking of DS_{lt} (that is, $SQ(r)$). If the answer is affirmative, the capacity supply is sufficient to meet the capacity requirement with the consideration of the types and amount of process capabilities of each machine. Otherwise, further adjustment of the maximum loading level is required to meet the cumulative capacity demand.

The maximum loading level among the machines obtained after the two-phase capacity supply assessment is set as a basis for setting the UBLU. Followings are the detail computation steps:

Capacity requirement of each process capability

Step 1: Calculate capacity requirement (DC_{lt}) of process capability in each planning period within the planning horizon.

$$DC_{lt} = \sum_i \sum_j (p_{ij} CR_{ijt} LT_{ijt}) \quad , \text{ for each } l, \text{ each } t \tag{14}$$

Step 2: Calculate average capacity loading (AL_l) of machines in photolithography workstation in each planning period.

$$AL_l = \sum_l DC_{lt} / K \quad , \text{ for each } t \tag{15}$$

Phase I. Independent capacity supply assessment

Step 1: Set $t = 1, l = 1$.

Step 2: Set $ML_l = AL_l$.

Step 3: Verify if independent capacity supply of process capability l is sufficient in planning period t . If yes, then go to step 5; else go to step 4.

$$DC_{lt} \leq SC_{lt} \tag{16}$$

where

$$SC_{lt} = ML_l \times \sum_k A_{kl}$$

Step 4: Adjust the maximum loading level (ML_t) to satisfy capacity requirement of process capability l .

$$ML_t = ML_t + (DC_{lt} - SC_{lt}) / \sum_k A_{kl} \tag{17}$$

Step 5: Check if $l = L$. If yes, then go to step 6; else let $l = l + 1$ and go to step 3.

Step 6: Check if $t = T$. If yes, then end of Phase I; else let $t = t + 1$, $l = 1$, and go to step 2.

Phase II. Dependent capacity supply assessment

Step 1: Set $t = 1$.

Step 2: Calculate the ratio (DS_{lt}) of each process capability.

$$DS_{lt} = DC_{lt} / (ML_t \times \sum_k A_{kl}) \text{ , for each } l \tag{18}$$

Step 3: Rank the values of all DS_{lt} in planning period t from large to small. Use r to represent the rank and $SQ(r)$ to represent the r -th process capability.

Step 4: Set $r = 1$.

Step 5: Calculate whether dependent capacity supply is sufficient. If equation (19) is satisfied, then go to step 7; else go to step 6.

$$\sum_{r=1}^r DC_{SQ(r),t} \leq CS_{rt} \tag{19}$$

where

$$CS_{rt} = ML_t \times \sum_k \min \left\{ 1, \max \left\{ \sum_{r=1}^r A_{k,SQ(r)}, 0 \right\} \right\}$$

Step 6: Adjust the maximum loading level (ML_t) to satisfy cumulative capacity requirement. Then go to step 7.

$$ML_t = ML_t + (\sum_{r=1}^r DC_{SQ(r),t} - CS_{rt}) / CS_{rt} \tag{20}$$

Step 7: Check if $r = L$. If yes, then go to step 8; else let $r = r + 1$ and go to step 5.

Step 8: Check if $t = T$. If yes, then end of Phase II; else let $t = t + 1$, and go to step 2.

Setting the upper bound of load unbalance (UBLU)

After the two-phase assessment above, the upper bound of load unbalance can be set as $(ML_t - AL_t) / AL_t$. However, considering the processing time for each layer of each product is not identical and is not divisible. It is revised as follows:

$$E_t = \max \left\{ (ML_t - AL_t) / AL_t \text{ , } \min_{\forall i,j} \{ p_{ij} \} / AL_t \right\} \tag{21}$$

3 Comparisons Among MILP, I-CSP, and O-CSP Models

A simple example is presented here to compare the performance between MILP, I-CSP, and O-CSP models. Consider a production environment with three machines, M1, M2 and M3, and each machine possesses different process capabilities as shown in Table 1.

Table 1. Process capability of machines

Machine No.	Process capability			
	1	2	3	4
1	1*	1	0	0
2	0	1	1	0
3	0	1	1	1

* 1 if the machine has this certain process capability; 0, otherwise.

Table 2. Processing time, loading occurrence time, process window constraint and critical operation of orders

Order No.	Layer Number						
	1	2	3	4	5	6	7
1	12,1,1,0*	15,1,3,1	19,2,2,0	12,2,3,1	14,3,2,0	-	-
2	11,1,1,0	16,1,3,1	18,2,2,0	11,2,3,1	9,3,2,0	15,3,3,1	17,3,1,0
3	13,1,2,0	15,1,3,0	10,2,4,1	13,2,2,0	20,3,4,1	12,3,3,0	-
4	12,1,2,0	14,2,4,1	14,2,2,0	13,2,4,1	12,3,3,0	15,3,2,0	-
5	13,1,2,0	13,1,3,0	19,1,4,1	12,2,3,0	13,2,4,1	12,3,2,1	16,3,2,0

* processing time (hr), load occurrence time (week), required process capability, whether a critical operation (1: critical operation; 0: non-critical operation), respectively.

Table 3. Performance of different solving models

Model		[k, t]								
		[1, 1]	[2, 1]	[3, 1]	[1, 2]	[2, 2]	[3, 2]	[1, 3]	[2, 3]	[3, 3]
MILP 0.0357 ¹ 0.28 ²	(1) u_{kt}^+	0.0099	0.0000	0.0000	0.0019	0.0000	0.0019	0.0000	0.0039	0.0000
	(2) u_{kt}^-	0.0000	0.0019	0.0079	0.0000	0.0039	0.0000	0.0019	0.0000	0.0019
	(3) ³	0.0099	0.0019	0.0079	0.0019	0.0039	0.0019	0.0019	0.0039	0.0019
	(4) ⁴		0.0099			0.0039			0.0039	
	(5) ⁵		1.6632			0.6552			0.6552	
I-CSP 2.7694 0.00	(1) u_{kt}^+	0.4623	0.0000	0.0000	0.3591	0.0000	0.0000	0.5634	0.0000	0.0000
	(2) u_{kt}^-	0.0000	0.1865	0.2757	0.0000	0.0634	0.2956	0.0000	0.2817	0.2817
	(3)	0.4623	0.1865	0.2757	0.3591	0.0634	0.2956	0.5634	0.2817	0.2817
	(4)		0.4623			0.3591			0.5634	
	(5)		77.6664			60.3288			94.6512	
O-CSP 0.0357 0.30	(1) u_{kt}^+	0.0099	0.0000	0.0000	0.0019	0.0000	0.0019	0.0000	0.0039	0.0000
	(2) u_{kt}^-	0.0000	0.0019	0.0079	0.0000	0.0039	0.0000	0.0019	0.0000	0.0019
	(3)	0.0099	0.0019	0.0079	0.0019	0.0039	0.0019	0.0019	0.0039	0.0019
	(4)		0.0099			0.0039			0.0039	
	(5)		1.6632			0.6552			0.6552	

¹ Objective function value. Notice the objective function value of I-CSP and O-CSP is the sum of ($u_{kt}^+ + u_{kt}^-$).
² Computational time (sec). Notice the time of O-CSP is the sum of solving time of the 15 iterations in Table 4.
^{3,4,5} (3)=(1)+(2). (4) is the maximum value of (3) under t. (5)=(4)×available capacity (168 hours/period).

There are five orders to be released, and the information of processing time (hrs), loading occurrence time (week), required process capability and critical layer process are shown in Table 2. The commercial software ILOG OPL 3.5 [6] is utilized to solve the simplified example by three different models: MILP model, I-CSP model and O-CSP model. The results are shown in Table 3.

Even though I-CSP model uses the least amount of solving time among the three models, its objective function value (2.7694) and workstation utilization rate difference (maximum difference of 0.5634) are the worst. As an extension of I-CSP model,

Table 4. Computation process of O-CSP in solving the simple example

# of iterations	UBLU	Objective function value	Solving time	Maximum difference ¹
1	-	2.7694	0.00	0.5634
2	0.5634	1.6782	0.02	0.3432
3	0.3432	1.3051	0.02	0.2599
⋮	⋮	⋮	⋮	⋮
6	0.2480	0.8924	0.03	0.1865
7	0.1865	0.6502	0.02	0.1448
8	0.1448	0.6146	0.02	0.0932
⋮	⋮	⋮	⋮	⋮
14	0.0277	0.1304	0.02	0.0218
15	0.0218	0.0357	0.05	0.0099
16	0.0099		no feasible solution	

¹The maximum value of $(u_{kt}^+ + u_{kt}^-)$.

O-CSP can constantly adjust the UBLU by using equation (13) (see Table 4 for computation process) to eventually derive at an optimal solution (same as the objective function value of MILP model). One drawback of the model is that the required iterations of adjustments are not estimable. With the upper bound of load unbalance estimation introduced in section 2.3, we could set the upper bound of load unbalance to 0.1811. Comparing with the data in Table 4, 0.1811 is the upper bound for the 7th iteration in the O-CSP model. In consequence, the adoption of the setting of UBLU to the O-CSP model can effectively reduce the number of iterations.

4 A Real-World Application

To verify the applicability of the proposed model, a real-world case investigated in [3], is examined here. In this wafer fab, there are ten steppers and five different process capabilities. Five types of products, A, B, C, D and E, are manufactured, and each product requires 17, 19, 16, 20 and 19 times of photolithography operations respectively. The total required photolithography operation time for a product is in a range between 597 to 723 minutes. Product A and B require fabrication technology of 0.17 μm, while Product C, D and E adopt 0.14 μm fabrication technology. Production planning and control department sets the planning horizon to be 28 days and planning period to be 7 days. In the planning horizon, there are 474 lots that are expected to be released. Manufacturing execution system (MES) reveals that there are currently 204 lots of WIP on floor.

The CAPP problem is solved by O-CSP using software ILOG OPL 3.5 [6], and the results are shown in Table 5. Through the upper bound of load unbalance estimation, the CAPP problem could have a fairly balanced capacity allocation result (i.e. $ML_t = AL_t$), and the UBLU is set to 0.0014 (=15/10800, where minimum processing time is 15 minutes and available capacity for machines in a planning period is 10,800 minutes.) Table 5 shows that the objective function value derived from the CAPP problem is 0.0205 and the required solving time is 475.22 sec. This result is superior than that generated by Chung *et al.* [3] that the objective function value and required solving time are 0.0291 and 5.3878 hours respectively.

Table 5. Objective function value, maximum difference and solving time under different UBLU

UBLU	Objective function value	Maximum difference ¹	Solving time (sec.)
—	9.4434	0.9994	125.94
⋮	⋮	⋮	⋮
0.0014	0.0205	0.0013	475.22
0.0013	0.0167	0.0010	576.53
0.0010	0.0141	0.0008	851.02
0.0008	0.0116	0.0007	903.43
0.0007	0.0107	0.0006	973.98
0.0006	0.0084	0.0005	3477.64
0.0005	no feasible solution		

¹ The maximum value of $(u_{it}^+ + u_{it}^-)$.

When the UBLU (0.0014) is used as a basis of O-CSP for solving the CAPP problem, the required number of iterations is only six times. This indicates that the setting of UBLU can effectively reduce the search space of O-CSP. Such a solving process possesses a very good quality and has its application value in real practice.

5 Conclusion

In this paper, we consider the capacity allocation problem with the process window and machine dedication constraints that are apparent in wafer fabrication. We model the CAPP problem as a constraint satisfaction problem (CSP), which uses an efficient search algorithm to obtain a feasible solution. A relatively large amount of setting and calculation process is required in CSP because it treats the objective function as one of the constraints for searching an optimal solution while the bound of objective function is narrowed down through an iterative process. Hence, we propose a method for setting the upper bound of load unbalance among machines, and the search space and the number of computations can be decreased effectively in the CAPP problem. The result shows that a very good solution can be obtained in a reasonable time and can be a reference for wafer lot release and dispatching of photolithography machines, and the model thus is valuable in real world application.

Acknowledgements

This paper is partially support by Grant NSC94-2213-E-009-086.

References

1. Akçali, E., Nemoto, K., Uzsoy, R.: Cycle-Time Improvements for Photolithography Process in Semiconductor Manufacturing. *IEEE Transactions on Semiconductor Manufacturing*. **14**(1) (2001) 48-56.
2. Brailsford, S.C., Potts, C.N., Smith, B.M.: Constraint Satisfaction Problems: Algorithms and Applications, *European Journal of Operational Research*. **119**(3) (1999) 557-581.

3. Chung, S.H., Huang, C.Y., Lee, A.H.I.: Capacity Allocation Model for Photolithography Workstation with the Constraints of Process Window and Machine Dedication. *Production Planning and Control*. Accepted. (2006).
4. Freuder, E.C., Wallace, R.J.: *Constraint Programming and Large Scale Discrete Optimization*. American Mathematical Society. (2001)
5. Hung, Y.F., Cheng, G.J.: Hybrid Capacity Modeling for Alternative Machine Types in Linear Programming Production Planning. *IIE Transactions*. **34**(2) (2002) 157-165.
6. ILOG Inc.: *ILOG OPL Studio 3.5*. ILOG Inc., France. (2001)
7. Kim, S., Yea, S.H., Kim, B.: Shift Scheduling for Steppers in the Semiconductor Wafer Fabrication Process. *IIE Transactions*. **34**(2) (2002) 167-177.
8. Kishimoto, M., Ozawa, K., Watanabe, K., Martin, D.: Optimized Operations by Extended X-Factor Theory Including Unit Hours Concept. *IEEE Transactions on Semiconductor Manufacturing*. **14**(3) (2001) 187-195.
9. Leachman, R.C., Carmon, T.F.: On Capacity Modeling for Production Planning with Alternative Machine Types. *IIE Transactions*. **24**(4) (1992) 62-72.
10. Lee, Y.H., Park, J., Kim, S.: Experimental Study on Input and Bottleneck Scheduling for a Semiconductor Fabrication Line. *IIE Transactions*. **34**(2) (2002) 179-190.
11. Lustig, I.J., Puget, J.-F.P.: Program Does Not Equal Program: Constraint Programming and Its Relationship to Mathematical Programming. *Interfaces*. **31**(6) (2001) 29-53.
12. Toktay, L.B., Uzsoy, R.: A Capacity Allocation Problem with Integer Side Constraints. *European Journal of Operational Research*. **109**(1) (1998) 170-182.
13. Uzsoy, R., Lee, C.-Y., Martin-Vega, L.A.: A Review of Production Planning and Scheduling Models in the Semiconductor Industry (I): System Characteristics, Performance Evaluation and Production Planning. *IIE Transactions*. **24**(4) (1992) 47-60.
14. Uzsoy, R., Lee, C.-Y., Martin-Vega, L.A.: A Review of Production Planning and Scheduling Models in the Semiconductor Industry (II): Shop-Floor Control. *IIE Transactions*. **26**(5) (1994) 44-55.

Scheduling an R&D Project with Quality-Dependent Time Slots

Mario Vanhoucke^{1,2}

¹ Faculty of Economics and Business Administration,
Ghent University, Gent, Belgium

² Operations & Technology Management Centre,
Vlerick Leuven Gent Management School, Gent, Belgium
mario.vanhoucke@ugent.be

Abstract. In this paper we introduce the concept of quality-dependent time slots in the project scheduling literature. Quality-dependent time slots refer to pre-defined time windows where certain activities can be executed under ideal circumstances (optimal level of quality). Outside these time windows, there is a loss of quality due to detrimental effects. The purpose is to select a quality-dependent time slot for each activity, resulting in a minimal loss of quality. The contribution of this paper is threefold. First, we show that an R&D project from the bio-technology sector can be transformed to a resource-constrained project scheduling problem (RCPSP). Secondly, we propose an exact search procedure for scheduling this project with the aforementioned quality restrictions. Finally, we test the performance of our procedure on a randomly generated problem set.

1 Introduction

In the last decades, the research in resource-constrained project scheduling has been investigated from different angles and under different assumptions. The main focus on project time minimization has shifted towards other objectives (e.g. net present value maximization), extensions such as multi-mode scheduling and/or preemption, and many other facets (for the most recent overview, see [1]).

However, the literature on project scheduling algorithms where quality considerations are taken into account is virtually void. [9] maximize the quality in the resource-constrained project scheduling problem by taking the rework time and rework cost into account. They argue that their work is a logical extension of the classical resource-constrained project scheduling efforts. Therefore, they refer to a previous study by the same authors that indicated that over 90% of the project managers take the maximization of the quality of projects and their outcomes as their primal objective. Given that emphasis on quality management and its implementation in project management, and the need to develop new tools and techniques for scheduling decisions, we elaborate on that issue based on a real-life project in the bio-technology sector.

The motivation of this research paper lies in the effort we put in the scheduling of a project with genetically manipulated plants. In this project, several activities need to be scheduled in the presence of limited resources and severe quality restrictions. More

precisely, some activities need to be executed preferably within certain pre-defined periods, referred to as *quality-dependent time slots*. Although the execution is also possible outside these pre-defined intervals, it is less desirably since it leads to a decrease in quality. The concept of pre-defined time windows for activity execution is not new in the project scheduling literature. [2] criticize the traditional models in which it is assumed that an activity can start at any time after the finishing of all its predecessors. To that purpose, they consider two improvements over the traditional activity networks by including two types of time constraints. *Time-window constraints* assume that an activity can only start within a specified time interval. *Time-schedule constraints* assume that an activity can only begin at one of an ordered schedule of beginning times. [15] elaborate on these time constraints and argue that time can be treated as a repeating cycle where each cycle consists of two categories: (i) some pairs of rest and work windows and (ii) a leading number specifying the maximal number of time each pair should iterate. By incorporating these so-called *time-switch constraints*, activities are forced to start in a specific time interval and to be down in some specified rest interval. *Quality-dependent time slots* refer to pre-defined time windows where certain activities can be executed under ideal circumstances (optimal level of quality). Outside these time windows, there is a loss of quality due to detrimental effects. Unlike the time-switch constraints, the quality-dependent time slots allow the execution of the activity outside the pre-defined window leading to an extra cost or decrease in quality. Consequently, the quality-dependent time slots are similar to the time-switch constraints. The latter are hard constraints (execution is only possible *within* the interval) and the former are soft constraints (execution is preferable *within* the interval but is also possible *outside* the interval) that can be violated at a certain penalty cost.

Our project settings assume that each activity has several pre-defined quality-dependent time slots, from which one has to be selected. The selection of a time slot must be done before the start of the project (in the planning phase). Given a fixed set of time slots per activity, the target is then to select a time slot and to schedule the project such that the loss in quality will be minimized.

2 Description of the Problem

The project under study can be represented by an activity-on-the-node network where the set of activity nodes, N , represents activities and the set of arcs, A , represents finish-start precedence constraints with a time lag of zero. The activities are numbered from the dummy start activity 1 to the dummy end activity n and are topologically ordered, i.e. each successor of an activity has a larger activity number than the activity itself. Each activity has a duration d_i ($1 \leq i \leq n$) and a number of quality-dependent time windows $nr(i)$. Each window l of activity i ($1 \leq i \leq n$ and $1 \leq l \leq nr(i)$) is characterized by a time-interval $\left[\overset{-}{q_{il}}, \overset{+}{q_{il}} \right]$ of equal quality, while deviations outside

that interval result in a loss of quality. Note that the time slot $\left[\overset{-}{q_{il}}, \overset{+}{q_{il}} \right]$ is used to refer to a window with optimal quality and can be either an interval or a single point-in-time. The quality deviation of each activity i can be computed as

$Q_i^{\text{loss}} = \max\{q_{il}^- - s_i ; s_i - q_{il}^+ ; 0\}$ and depends on the selection of the time window l , with s_i the starting time of activity i . To that purpose, we need to introduce a binary decision variable in our conceptual model which determines the selection of a specific time interval for each activity i , $y_{il} = \begin{cases} 1, & \text{if time interval } l \text{ has been selected for activity } i \\ 0, & \text{otherwise} \end{cases}$.

We use q_{il}^{opt} to denote the minimal activity cost associated with a fixed and optimal level of quality for each time window l of activity i . We use q_{il}^{extra} to denote the loss in quality per time unit deviation from the time interval and consequently, the total cost of quality equals $\sum_{i=1}^n \sum_{l=1}^{nr(i)} (q_{il}^{\text{opt}} + q_{il}^{\text{extra}} Q_i^{\text{loss}}) y_{il}$. Note that $nr(0) = nr(n) = 1$, since

nodes 0 and n are dummy activities with $q_{01}^- = q_{01}^+$ and $q_{n1}^- = q_{n1}^+$. Moreover, we set $q_{01}^{\text{extra}} = \infty$ to force the dummy start activity to start at time instance zero. The project needs to be finished before a negotiated project deadline δ_n , i.e. $q_{n1}^- = q_{n1}^+ = \delta_n$. Consequently, setting $q_{n1}^{\text{extra}} = \infty$ denotes that the project deadline can not be exceeded (a hard constraint), while $q_{n1}^{\text{extra}} < \infty$ means that the project deadline can be exceeded at a certain penalty cost (soft constraint).

There are K renewable resources with a_k ($1 \leq k \leq K$) as the availability of resource type k and with r_{ik} ($1 \leq i \leq n, 1 \leq k \leq K$) as the resource requirements of activity i with respect to resource type k . The project with renewable resources and quality-dependent time windows can be conceptually formulated as follows:

$$\text{Minimize } \sum_{i=1}^n \sum_{l=1}^{nr(i)} (q_{il}^{\text{opt}} + q_{il}^{\text{extra}} Q_i^{\text{loss}}) y_{il} \tag{1}$$

Subject to

$$s_i + d_i \leq s_j \quad \forall (i, j) \in A \tag{2}$$

$$\sum_{i \in S(t)} r_{ik} \leq a_k \quad k = 1, \dots, K \text{ and } t = 1, \dots, T \tag{3}$$

$$Q_i^{\text{loss}} \geq \sum_{l=1}^{nr(i)} q_{il}^- y_{il} - s_i \quad i = 1, \dots, n \tag{4}$$

$$Q_i^{\text{loss}} \geq s_i - \sum_{l=1}^{nr(i)} q_{il}^+ y_{il} \quad i = 1, \dots, n \tag{5}$$

$$\sum_{l=1}^{nr(i)} y_{il} = 1 \quad i = 1, \dots, n \tag{6}$$

$$s_1 = 0 \quad (7)$$

$$s_i \in \text{int}^+, Q_i^{\text{loss}} \in \text{int}^+ \quad i = 1, \dots, n \quad (8)$$

$$y_{il} \in \text{bin} \quad i = 1, \dots, n \text{ and } l = 1, \dots, nr(i) \quad (9)$$

where $S(t)$ denotes the set of activities in progress in period $[t-1, t]$. The objective in Eq. 1 minimizes the total quality cost of the project (i.e. the fixed cost within the selected time window plus the extra cost of quality loss due to deviations from that interval). The constraint set given in Eq. 2 maintains the finish-start precedence relations among the activities. Eq. 3 represents the renewable resource constraints and the constraint sets in Eq. 4 and Eq. 5 compute the deviation from the selected time window of each activity. Eq. 6 represents the time window selection and forces to select a single time window for each activity. Eq. 7 forces the dummy start activity to start at time zero and Eq. 8 ensures that the activity starting times as well as the time window deviations assume nonnegative integer values. Eq. 9 ensures that the time window selection variable is a binary (0/1) variable. Remark that the quality loss function measuring the quality decrease due to a deviation from the ideal time window l can be off any form (such as stepwise functions, convex functions, etc...). However, we assume in Eqs. [1]-[9], without loss of generality, a linear quality deviation function.

Although our first acquaintance with this problem type was during the scheduling of a genetically manipulated plants project, we believe that there are numerous other examples where pre-defined time-windows need to be selected before the execution of the project. The following four examples illustrate the possible generalization of multiple quality-dependent time windows to other project environments:

Perishable Items. The project of this paper, which motivates us to elaborate on this issue, is a typical example where items (i.e. plants) are perishable. Many project activities consist of tests on growing plants where the quality is time-dependent since there is an optimal time interval of consumption. Earlier consumption is possible, at a cost of a loss in quality, since the plants are still in their ripening process. Later consumption results in loss of quality due to detrimental effects.

State-of-Nature Dependencies. In many projects, the performance of some activities might depend on the state-of-nature. In this case, a pre-defined set of possible starting times depending on the state-of-nature are linked with possible execution times of the activity, and the deviation from these time windows is less desirable (resulting in higher costs or quality loss) or even completely intolerable.

Multiple Activity Milestones. The project scheduling literature with due dates (milestones) has been restricted to considering projects with pre-assigned due dates (see e.g. [11] and [12]). In reality, milestones are the results of negotiations, rather than simply dictated by the client of the project. Therefore, we advocate that due dates, including earliness and tardiness penalty costs for possible deviations, are agreed upon by the client and the contractor (and possibly some subcontractors). This results in a set of possible due dates for each activity, rather than a single pre-defined due date. The objective is then to select a due date for each activity such that the total earliness/tardiness penalty costs will be minimized.

Time-Dependent Resource Cost. In many projects, the cost of (renewable) resources heavily depends on the time of usage. The aforementioned time-switch constraints are a typical and extreme example of time-dependent resource costs, since it restricts the execution of activities to pre-defined time intervals (work periods) without any possibility to deviate. However, if we allow the activities to deviate from their original work periods (e.g. by adding more (expensive) resources to an activity in the pre-defined rest period), the work periods can be considered as the quality-dependent time slots while the rest periods are periods outside these slots in which the activity can be executed at an additional cost.

3 The Algorithm

The problem type under study requires the selection of a quality dependent time-window from a possible set of windows such that the total quality loss is minimized. A closer look to the problem formulation of (1)-(9) reveals the following observations:

- When $K = 0$, i.e. there are no resources with limited capacity, the problem given in (1)-(9) reduces to an unconstrained project scheduling problem with non-regular measures of performance. Due to the special structure of the quality loss functions and the multiple time-windows, the problem reduces to a separable nonconvex programming problem (see “solution algorithm for problem (1)-(2), (4)-(9)” described below).
- When $K > 0$, i.e. there is at least one renewable resource with limited capacity, resource conflicts can arise during the scheduling of the project. Therefore, this problem type can be solved to optimality by any branch-and-bound enumeration scheme for project scheduling problems with non-regular measures of performance (see “solution algorithm for problem (1)-(9)” described below).

In this section we describe a double branch-and-bound algorithm for the problem type under study. The first branch-and-bound procedure ignores the renewable resource constraints and searches for an exact solution of the unconstrained project scheduling problem. The second branch-and-bound procedure aims at resolving resource conflicts and needs the previously mentioned solution as an input in every node of the tree.

Solution Algorithm for Problem (1)-(2), (4)-(9): The basic idea of this solution approach relies on the approach used by [6] and [7]. Their problem is to find the vector $x = (x_1, \dots, x_n)$ which minimizes

$$\varphi(x) = \sum_{i=1}^n \varphi_i(x_i) \quad \text{subject to } x \in G \text{ and } l \leq x \leq L \quad (10)$$

for which it is assumed that G is closed and that each φ_i is lower semi-continuous, possibly nonconvex, on the interval $[l_i, L_i]$. In their paper, they have presented an algorithm for separable nonconvex programming problems. To that purpose, they solve a sequence of problems in a branch-and-bound approach in which the objective

function is convex. These problems correspond to successive partitions of the feasible set. This approach has been successfully applied for different optimization problems.

[10] have shown that this problem type is a special case of irregular starting-time costs project scheduling which can be formulated as a maximum-flow problem and hence, can be solved using any maximum-flow algorithm.

Due to the special structure of the convex envelope, we rely on an adapted procedure of [14] developed for an unconstrained project scheduling problem with activity-based cash flows which depend on the time of occurrence. This branch-and-bound procedure basically runs as follows. At each node, we calculate a lower bound for the total quality cost $q_{il}^{opt} + q_{il}^- Q_i^- + q_{il}^+ Q_i^+$ for each activity i . To that purpose, we construct the *convex envelope* of the total quality cost profile over the whole time window $[es_i, lf_i]$ for each activity i (es_i = earliest start and lf_i = latest finish). The convex envelope of a function $F = q_{il}^{opt} + q_{il}^- Q_i^- + q_{il}^+ Q_i^+$ ($l = 1, \dots, nr(i)$) taken over $C = [es_i, lf_i]$ is defined as the highest convex function which fits below F .

If the reported solution is not feasible for the original problem, the algorithm starts to branch. The algorithm calculates two new convex envelopes for these subsets and solves two new problems at each node. Branching continues from the node with the lowest lower bound. If all activities at a particular node of the branch-and-bound tree are feasible, then we update the upper bound of the project (initially set to ∞) and explore the second node at that level of the branch-and-bound tree. Backtracking occurs when the calculated lower bound is larger than or equal to the current lower bound. The algorithm stops when we backtrack to the initial level of the branch-and-bound tree and reports the optimal lower bound. This lower bound can be calculated at each node of the branch-and-bound algorithm described hereunder.

Solution Algorithm for Problem (1)-(9): In order to take the renewable resource constraints into account (i.e. equation (3)) we rely on a classical branch-and-bound approach that uses the unconstrained solutions as lower bounds at each node of the tree. Since the previous branch-and-bound procedure searches for an optimal solution for the unconstrained project and consequently, ignores the limited availability of the renewable resources, the second branching strategy boils down to resolving resource conflicts. If resource conflicts occur, we need to branch. A resource conflict occurs when there is at least one period $]t - 1, t]$ for which $\exists k \leq K : \sum_{i \in S(t)} r_{ik} > a_k$. To that

purpose, we rely on the branch-and-bound approach of [8] for the resource-constrained project scheduling problem with discounted cash flows. This is an adapted version of the branching scheme developed by [3] for the resource-constrained project scheduling problem and is further enhanced by [5], [12] and [13].

In order to prune certain nodes of the branch-and-bound tree, we have implemented the so-called *subset dominance rule*. This dominance rule has originally been developed by [5] and has been applied in the branch-and-bound procedures of [12]

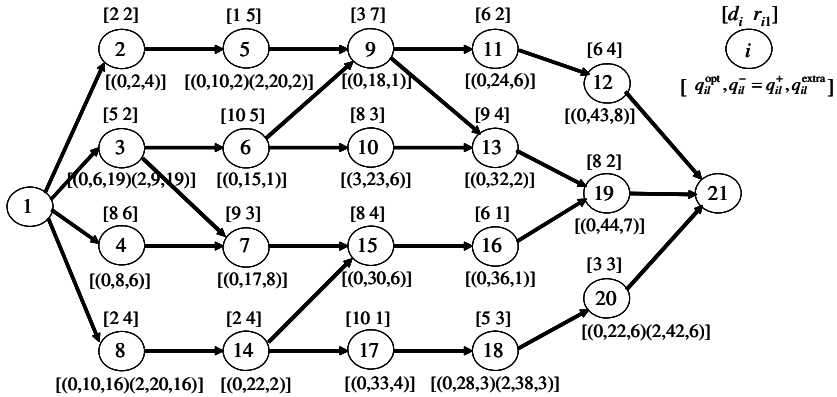


Fig. 1. An example project with quality-dependent time slots

and [13]. This dominance rule can be applied when the set of added precedence constraints (to resolve resource conflicts) of a previously examined node in the tree is a subset of the set of precedence constraints of the current node.

Example: We illustrate the project scheduling problem with limited resources and quality-dependent time slots by means of an example project of figure 1. The two numbers above each node are used to denote the activity duration d_i and its requirement r_{i1} for a single renewable resource with availability $a_1 = 10$. The numbers below the node are used to denote $(q_{il}^{opt}, q_{il}^- = q_{il}^+, q_{il}^{extra})$ for each quality-dependent time slot l . For the sake of clarity, we assume that $q_{il}^- = q_{il}^+$, i.e. the quality-dependent time slots are a single point in time. Moreover, we assume q_{il}^{extra} is equal for each interval l .

Each activity of the example project belongs to one of the following categories. An activity can be the subject to single (activities 12 and 17) or multiple quality-dependent time slots (activities 3, 5, 8, 18 and 20). Activities can also have the requirement to be scheduled as-soon-as-possible (ASAP; activities 2, 4, 6, 7, 9, 10, 11 and 13) or as-late-as-possible (ALAP; activities 14, 15, 16, 19 and 21). This can be incorporated in the network by adding a single quality-dependent time slot with $q_{il}^- = q_{il}^+ = es_i$ (ASAP) or $q_{il}^- = q_{il}^+ = ls_i$ (ALAP), with es_i (ls_i) the earliest start (finishing) time of activity i . Deviations from these requirements will be penalized by q_{il}^{extra} per time unit. We assume that the project deadline T equals 44 time units.

Figure 2 displays the schedules found by solving the RCPSP without (i.e. minimization of project time) and with the quality-dependent time slots. The activities highlighted in dark grey are the activities that are scheduled at a different time instance between the two schedules. Activities 3, 6, 8, 14, 10, 17, 13, 18, 12 and 20 have been scheduled later than the classical RCPSP schedule, while activities 5 and 7 have been scheduled earlier.

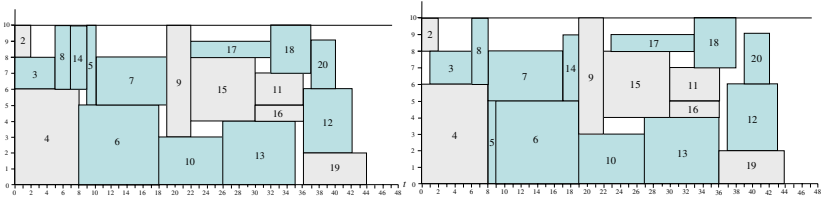


Fig. 2. The RCPSP schedule with minimal time (left) and quality-dependent time slots (right)

4 Computational Experience

In order to validate the efficiency, we have coded our double B&B procedure in Visual C++ Version 6.0 under Windows NT 4.0 on a Compaq personal computer (Pentium 500 MHz processor). We have generated a problem set by the RanGen network generator of [4] with 10, 20 and 30 activities.

In order to test the presence of renewable resources on the performance of our B&B procedure, we have extended each network instance with renewable resources under a pre-defined design. To that purpose, we rely on the resource use *RU* and the resource-constrainedness *RC* which can be generated by RanGen. More precisely, we use 4 settings for the *RU* (1, 2, 3 or 4) and 4 settings for the *RC* (0.25, 0.50, 0.75 or 1).

In order to generate data for the quality-dependent time slots, we need to generate values for the number of time slots $nr(i)$, the start and finishing time instance per time slot l (\bar{q}_{il}^- and \bar{q}_{il}^+) and the quality deviation per time unit for each time slot l (q_{il}^{extra}). We used 4 different settings to generate the number of time slots per activity, i.e. $nr(i)$ equals 5, 10, 15 or 20. The start and finishing times of each time slot have been carefully selected between the earliest start time and the latest finish time of each activity, such that the ‘spread’ of the time slots have been generated under 3 settings, i.e. low (time slots close to each other), average and high (time slots far from each other). The q_{il}^{extra} values have been randomly generated. Using 30 instances for each setting, we have created 17,280 problem instances.

In table 1 we report the computational results for the project scheduling problem with renewable resource constraints and quality-dependent time-slots. To that purpose, we display the average CPU-time in seconds (Avg.CPU), the number of problems solved to optimality within 100 seconds CPU-time (#Solved), the average number of created nodes in the main branch-and-bound tree (#Avg.CN) to resolve resource conflicts, the average number of branched nodes for this B&B tree (Avg.BN) and the average number of created nodes (Avg.CN2) in the unconstrained branch-and-bound tree (lower bound calculation) at each node of the main B&B tree. The row labelled ‘all instances’ gives the average results over all 17,280 problem instances and illustrates the efficiency of our double branch-and-bound procedure. In the remaining rows we show more detailed results for the different parameters of our full factorial experiment.

Table 1. Computational results for the RCPSP with quality-dependent time slots

		Avg.CPU	#Solved	Avg.CN	Avg.BN	Avg.CN2
All instances		43.517	10,212	78.924	71.517	22.881
Number of activities	10	0.099	5,760	57.378	44.173	7.825
	20	53.149	3,012	86.021	80.557	26.672
	30	77.304	1,440	93.372	89.821	34.145
RU	1	30.462	3,123	59.043	49.905	28.101
	2	43.544	2,530	80.558	72.422	24.646
	3	48.318	2,348	87.304	80.869	20.121
	4	51.745	2,211	88.791	82.872	18.654
RC	0.25	24.602	3,328	49.583	36.353	25.489
	0.5	48.772	2,338	85.655	79.643	25.258
	0.75	50.547	2,270	89.723	84.356	20.758
	1	50.149	2,276	90.735	85.716	20.017
nr(i)	5	31.851	3,069	76.684	70.006	10.714
	10	45.553	2,476	79.643	72.077	20.211
	15	48.146	2,339	79.423	71.860	28.533
	20	48.519	2,328	79.947	72.125	32.064
Spread	low	39.394	3,640	78.472	71.500	15.385
	mid	44.644	3,333	79.704	72.422	24.841
	high	46.515	3,239	78.596	70.629	28.416

As expected, the *RU* and the *RC* are positively correlated with the problem complexity. Indeed, the more resources in the project and the tighter their constrainedness, the higher the probability for a resource conflict to occur. These effects are completely in line with literature (see e.g. [13]). The number of time slots is positively correlated with problem complexity. The spread of the time slots is positively correlated with the problem complexity. When time slots are close to each other, the selection of an activity time slot is rather straightforward since only one (or a few) are relevant.

5 Conclusions and Areas for Future Research

In this paper we presented a double branch-and-bound procedure for the resource-constrained project scheduling problem with quality-dependent time slots. The depth-first branch-and-bound strategy to solve renewable resource conflicts makes use of another branch-and-bound procedure to calculate the lower bounds at each node. The branching scheme has been extended with the subset dominance rule to prune the search tree considerably. The incorporation of quality-dependent time slots in project scheduling is, to the best of our knowledge, completely new.

It is in our future intentions to broaden the research efforts of quality-dependent time slot selection in project scheduling. More precisely, the introduction of *dynamic quality-dependent time slots* should open the door to many other applications. In this case, each activity can be executed several times, preferably within the pre-defined time slots. Moreover, the different time slots per activity depend on each other, in the sense that the time interval $\left[q_{il}^-, q_{il}^+ \right]$ depends on the finishing time of the activity in or around the previous time slot $= \left[q_{il-1}^-, q_{il-1}^+ \right]$. A typical example is a maintenance

operation that needs to be done within certain time limits. A second maintenance operation depends on the execution of the first one, resulting in a second time slot that depends on the starting time of the activity around the first time slot. The development of heuristic solution methods to solve larger, real-life problems instances also lies within our future research intentions.

References

1. Brücker, P., Drexl, A., Möhring, R., Neumann, K., Pesch, E.: Resource-constrained project scheduling: notation, classification, models and methods. *European Journal of Operational Research*. 112 (1999) 3-41
2. Chen, Y.L., Rinks, D., Tang, K.: Critical path in an activity network with time constraints. *European Journal of Operational Research*. 100 (1997) 122-133
3. Demeulemeester E., Herroelen, W.: A branch-and-bound procedure for the multiple resource-constrained project scheduling problem. *Management Science*. 38 (1992) 1803-1818
4. Demeulemeester, E., Vanhoucke, M., Herroelen, W.: A random network generator for activity-on-the-node networks. *Journal of Scheduling*. 6 (2003) 13-34
5. De Reyck, B., Herroelen, W.: An optimal procedure for the resource-constrained project scheduling problem with discounted cash flows and generalized precedence relations. *Computers and Operations Research*. 25 (1998) 1-17
6. Falk, J.E., Soland, R.M.: An algorithm for separable nonconvex programming problems. *Management Science*. 15 (1969) 550-569
7. Horst, R.: Deterministic methods in constrained global optimization: Some recent advances and new fields of application. *Naval Research Logistics*. 37 (1990) 433-471
8. Icmeli, O., Erengüç, S.S.: A branch-and-bound procedure for the resource-constrained project scheduling problem with discounted cash flows. *Management Science*. 42 (1996) 1395-1408
9. Icmeli-Tukel, O., Rom, W.O.: Ensuring quality in resource-constrained project scheduling. *European Journal of Operational Research*. 103 (1997) 483-496
10. Möhring, R.H., Schulz, A.S., Stork, F., Uetz, M.: On project scheduling with irregular starting time costs. *Operations Research Letters*. 28 (2001) 149-154
11. Schwindt, C.: Minimizing earliness-tardiness costs of resource-constrained projects. In: Inderfurth, K., Schwoedjauer, G., Domschke, W., Juhnke, F., Kleinschmidt, P., Waescher, G. (eds.). *Operations Research Proceedings*. Springer. (1999) 402-407
12. Vanhoucke, M., Demeulemeester, E., Herroelen, W.: An exact procedure for the resource-constrained weighted earliness-tardiness project scheduling problem. *Annals of Operations Research*. 102 (2000) 179-196
13. Vanhoucke, M., Demeulemeester, E., Herroelen, W.: On maximizing the net present value of a project under renewable resource constraints. *Management Science*. 47 (2001) 1113-1121
14. Vanhoucke, M., Demeulemeester, E., Herroelen, W.: Progress payments in project scheduling problems. *European Journal of Operational Research*. 148 (2003) 604-620
15. Yang, H.H., Chen, Y.L.: Finding the critical path in an activity network with time-switch constraints. *European Journal of Operational Research*. 120 (2000) 603-613

The Bottleneck Tree Alignment Problems

Yen Hung Chen^{1,*} and Chuan Yi Tang²

¹ Department of Applied Mathematics,
National University of Kaohsiung, Kaohsiung 811, Taiwan, R.O.C.
dr884336@cs.nthu.edu.tw

² Department of Computer Science,
National Tsing Hua University, Hsinchu 300, Taiwan, R.O.C.
cytang@cs.nthu.edu.tw

Abstract. Given a set W of k sequences (strings) and a tree structure T with k leaves, each of which is labeled with a unique sequence in W , a *tree alignment* is to label a sequence to each internal node of T . The weight of an edge of the tree alignment is the distance, such as *Hamming* distance, *Levenshtein (edit)* distance or *reversal* distance, between the two sequences labeled to the two ends of the edge. The *bottleneck tree alignment problem* is to find a tree alignment such that the weight of the largest edge is minimized. A *lifted tree alignment* is a tree alignment, where each internal node v is labeled one of the sequences that was labeled to the children of v . The *bottleneck lifted tree alignment problem* is to find a lifted tree alignment such that the weight of the largest edge is minimized. In this paper, we show that the *bottleneck tree alignment problem* is NP-complete even when the tree structure is the binary tree and the weight function is *metric*. For special cases, we present an exact algorithm to solve the *bottleneck lifted tree alignment problem* in polynomial time. If the weight function is *ultrametric*, we show that any *lifted tree alignment* is an optimal *bottleneck tree alignment*.

Keywords: Edit distance, bottleneck tree alignment, metric, ultrametric, NP-complete.

1 Introduction

Tree alignment [3, 9] is one of the fundamental problems in computational biology. Given a set W of k sequences (as DNA sequences of species) and a tree structure T with k leaves, each of which is labeled with a unique sequence, the *tree alignment* is to label a sequence to each internal node of T . The weight function of an edge of the tree is defined the distance, such as *Hamming* distance, *Levenshtein (edit)* distance [1] or *reversal* distance [7], between the two sequences labeled to the two ends of the edge. The *tree alignment problem* (TAP for short) is to find a tree alignment such that the sum of the weights of all its edges is minimized. This problem was showed to be NP-complete [12] even when the tree structure is a binary tree and some polynomial time approximation

* Corresponding author.

schemes (PTASs) had been proposed [8, 13, 14, 15]. When the tree structure is a star (also called *star alignment* or *median string*), this problem was also showed to be NP-complete [6, 9] when the weight function is *metric* (i.e., the weights of edges satisfy the triangle inequality). This problem was also showed to be MAX SNP-hard [12], but the weight function is not metric.

Ravi and Kececioğlu [8] posted a variant of the TAP with bottleneck objective function. It is the *bottleneck tree alignment problem* (BTAP for short). This problem is to find a tree alignment such that the weight of the largest edge is minimized. They gave an $O(\log k)$ -approximation algorithm in linear time. However, whether this problem is NP-complete still open. A *lifted tree alignment* [5, 13, 14] is a tree alignment, in which each internal node v is labeled one of the sequences that was labeled to the children of v . The *lifted tree alignment problem* is to find a lifted tree alignment such that the sum of the weights of all its edges is minimum and a polynomial time algorithm had been proposed [5]. An optimal solution of the lifted tree alignment problem is a good approximation solution for the TAP [8, 13, 14, 15]. The *bottleneck lifted tree alignment problem* (BLTAP for short) is to find a lifted tree alignment such that the weight of the largest edge is minimized. In this paper, we use $d(u, v)$ to denote the weight (simultaneously, distance) of the two ends (simultaneously, two sequences) u and v of an edge.

The rest of this paper is organized as follows. In section 2, we show that the BTAP is NP-complete even when the tree structure is the binary tree and the weight function is metric. In Section 3, we give an $O(k^2 n^2 + k^3)$ time algorithm to optimally solve the BLTAP, where n is the maximum length of sequences in W . In section 4, we show that any *lifted tree alignment* is an optimal solution of the BTAP when the weight function is *ultrametric*. The ultrametric [2, 5, 11, 17] is the metric [2, 5, 11] and satisfies the following condition: for any three sequences x , y , and z ,

$$\max\{d(x, y), d(y, z)\} \geq d(x, z).$$

Finally, we make a conclusion in section 5.

2 The Bottleneck Tree Alignment Problem Is NP-Complete

In this section, the distance function of two sequences is the edit distance which is metric. Edit distance has been thoroughly studied (see [1] for instance). The edit distance of the two sequences u and v is defined as the minimum number operations such that u transforms into v , where an operation is a single-symbol deletion, a single-symbol insertion, or a single-symbol substitution. Edit distance of the two sequences u and v can be computed in $O(|u| * |v|)$ time via dynamic programming [10, 16], where $|u|$ and $|v|$ are lengths of u and v .

Lemma 1. [6] *Given two sequences u and v ; if $|u| \geq |v|$ then $d(u, v) \geq |u| - |v|$.*

Lemma 2. [6] *Given two sequences u and v ; if u and v have no symbol in common then $d(u, v) = \max(|u|, |v|)$.*

A sequence is a string over some alphabet Σ and the set of all finite-length sequences of symbols from Σ is denote by Σ^* . Now, we definition the decision version of the BTAP as follows.

Bottleneck Tree Alignment Problem With Decision Version

Instance: A positive integer m , a set W of k distinct sequences $\{w_1, w_2, \dots, w_k\}$ over an alphabet Σ and a tree structure T with k leaves, each of which is labeled with a unique sequence in W .

Question: Does there exist sequences in Σ^* which are labeled to all internal nodes of T such that the weight of the largest edge in T is less than or equal to m , where the weight of an edge is the (edit) distance between the two sequences labeled to the two ends of the edge ?

Given a finite set S of sequences over an alphabet Σ , the *longest common subsequence problem* [4] is to find a longest length sequence s such that s is a subsequence of each sequence in S . Bonizzoni and Vedova [6] defined the *longest common subsequence0 (LCS0) problem*. Given a positive integer m and a finite set S of k distinct sequences over an alphabet Σ_0 , in which the length of each sequence in S is $2m$, the LCS0 problem is to find a common subsequence s of each sequence in S such that $|s| \geq m$. This problem was shown to be NP-complete [6].

LCS0 problem

Instance: A positive integer m and a set S of k distinct sequences s_1, s_2, \dots, s_k over an alphabet Σ_0 , in which the length of each sequence is $2m$.

Question: Does there exist a sequence s with $|s| \geq m$ such that s is a common subsequence of each sequence $s_i \in S$?

Theorem 1. *The bottleneck tree alignment problem is NP-complete even when the tree structure is a binary tree and the weight function is metric.*

Proof. First, It is not hard to see the BTAP in NP. Now, we show the *reduction*: the transformation from the LCS0 problem to the BTAP. Let a set of sequences $S = \{s_1, s_2, \dots, s_k\}$ over an alphabet Σ_0 and a positive constant m be the instance of LCS0. The length of each sequence in S is $2m$. Now, we construct a binary tree T as in Figure 1a, where each T_i is a subtree shown in Figure 1b. Then, we find two different symbols b and c that do not belong to Σ_0 and an alphabet $\Sigma_a = \{a_1, a_2, \dots, a_k\}$ has k different symbols that do not belong to $\Sigma_0 \cup \{b, c\}$. Each leaf in T is labeled by a sequence over $\Sigma = \Sigma_0 \cup \Sigma_a \cup \{b, c\}$. Let sequences b^{2m} and c^{2m} be b and c repeating $2m$ times, respectively. The leaves $\{y_1, y_2, \dots, y_k, z_1, z_2, \dots, z_k\}$ of T are labeled as follows. If i is odd, y_i and z_i are labeled by $s_i b^{2m}$ and $a_i b^{2m}$, respectively. Otherwise, y_i and z_i are labeled by $s_i c^{2m}$ and $a_i c^{2m}$, respectively. We will show that there is a common subsequence s of S with $|s| = m$ if and only if there is a tree alignment of T such that the weight of the largest edge is m .

(Only if) Assume that there is a common subsequence s of each sequence in S with $|s| = m$. We label the sequences $s b^{2m}$ and $s b^m$ to x_i and u_i if i is odd, respectively. We also label the sequences $s c^{2m}$ and $s c^m$ to x_i and u_i if i is even, respectively. By lemma 2, the weight of the largest edge of T is m .

(If) Suppose that there exists sequences which are labeled to the internal nodes of T such that the weight of the largest edge is m . For each $1 \leq i \leq k$, let t_i be the sequence which is labeled to x_i . By lemma 2, we have $d(y_i, z_i) = 2m$. Hence, $d(y_i, x_i)$ and $d(x_i, z_i)$ can not be less than m simultaneously via triangle inequality. However, the weight of the largest edge in T is equal to m , we have $d(y_i, x_i) = m$ and $d(x_i, z_i) = m$. Hence, t_i must contain b^{2m} (respectively, c^{2m}) and can not contain the symbol a_i if i is odd (respectively, even). Now t_i can be written as $t'_i b^{2m}$ (respectively, $t'_i c^{2m}$) if i is odd (respectively, even). Moreover, t'_i must be a subsequence of s_i with $|t'_i| = m$ by lemma 1 and lemma 2. Furthermore, each sequence which is labeled to u_i must contain b^m (respectively, c^m) if i is odd (respectively, even), otherwise $d(x_i, u_i) > m$. Because $b \neq c$, for $1 \leq i < k$, we have each edge (u_i, u_{i+1}) has weight greater than or equal to m by lemma 2. However, $d(u_i, u_{i+1})$ can not exceed m , thus each sequence which is labeled to u_i must be $t'_i b^m$ (respectively, $t'_i c^m$) if i is odd (respectively, even). Moreover, $t'_i = s_i$, for $1 \leq i \leq k$. Therefore, s must be a common subsequence of each sequence in S and the length of s is m . \square

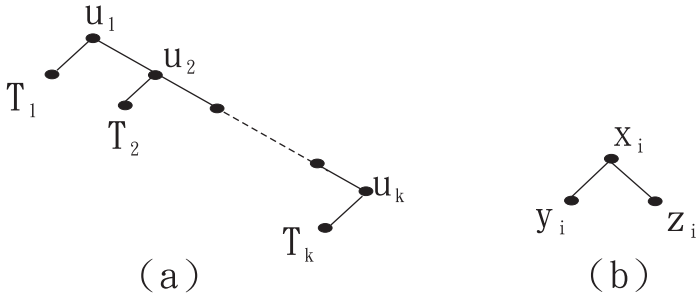


Fig. 1. a. The tree T . b. The subtree T_i .

3 The Bottleneck Lifted Tree Alignment Problem

In this section, we will present an exact algorithm for solving the BLTAP via dynamic programming. The distance function of two sequences also uses the edit distance. From now on, we let the tree T rooted at r and a set of sequences $W = \{w_1, w_2, \dots, w_k\}$ (i.e., labeling sequences of the leaves of T) be the instance of the BLTAP. First, let the level L_j of T be the nodes set, in which every node in L_j has the same height j in T and let H be the height of T . Hence, L_H is the set of all leaves of T and L_1 contains only root r . Then, for a node u in T , we let T_u be the subtree of T rooted at u and a sequences set $des(u)$ denote the labeling sequences of leaves of T_u . We also let the nodes set $chi(u)$ denote the children of u in T_u . For each node u and $w_i \in W$, let $B[u, w_i]$ denote minimum weight among the largest edges of all possible lifted tree alignments of T_u such that u is labeled by the sequence w_i . For convenience, we let the node v_α be one of the children of u with $w_i \in des(v_\alpha)$. For each $u \in L_j$, $1 \leq j \leq H$, if we have

all $B[v, w_t]$, in which each $v \in L_{j+1}$ is a child of u and $w_t \in des(v)$, then $B[u, w_i]$ can be computed as follow:

$$B[u, w_i] = \max\{B[v_\alpha, w_i], \max_{v \in \text{chi}(u) \text{ and } v \neq v_\alpha} \{ \min_{w_t \in des(v)} \{ \max(d(w_i, w_t), B[v, w_t]) \} \} \}. \tag{1}$$

It is clear that the weight of the largest edge for the optimal bottleneck lifted tree alignment is $\min_{w_i \in W} B[r, w_i]$. One can then obtain $B[r, w_i]$ by iteration above process for each $B[u, w]$ of a node u with $w \in des(u)$ via bottom-up (from leaves level to root). Then, it is not hard to find the optimal solution for the BLTAP via backtracking. For technical reason, for each leaf v , $B[v, w_i] = 0$ if the sequence w_i is a labeling sequence of v .

For clarification, we list the exact algorithm for the BLTAP as follows.

Algorithm Exact-BLTAP

Input: A tree T rooted at r with k leaves, each of which is labeled by k distinct sequences $W = \{w_1, w_2, \dots, w_k\}$.

Output: A lifted tree alignment of T .

1. **For** each sequence pair $(w_i, w_j) \in W$, compute the edit distance of w_i and w_j .
2. **For** each $u \in L_H$, let $B[u, w] = 0$ if the sequence w is a labeling sequence of u .
3. **/* recursively compute $B[u, w_i]$ */**
For $j = H - 1$ to 1 **do**
 for each node $u \in L_j$ with $w \in des(v)$, compute $B[u, w]$ by equation (1).
4. Select a sequence $w \in W$ such that the $B[r, w]$ is minimized among all sequences in W .
5. **/* backtracking */**
 Backtrack to find the optimal solution of the BLTAP.

Now, we analyze the time-complexity of this algorithm: step 1 takes $O(k^2n^2)$ time and step 2 takes $O(k)$ time. Computing each $B[u, w_i]$ with $w_i \in des(u)$ takes $O(k)$ time and then we need $O(k^2)$ time for each level. Hence, step 3 runs in $O(k^3)$ time. Clearly, step 4 and step 5 run in $O(k)$ time. The total time-complexity of Algorithm Exact-BLTAP is $O(k^2n^2 + k^3)$.

Theorem 2. *Algorithm Exact-BLTAP is an $O(k^2n^2 + k^3)$ time algorithm for solving the BLTAP.*

4 The Score Is Ultrametric

In this section, we show any lifted tree alignment is an optimal solution of the BTAP when the weight function d is ultrametric. Similarly, the tree T and a set of sequences $W = \{w_1, w_2, \dots, w_k\}$ (i.e., labeling sequences of the leaves of T) be the instance of the BTAP.

Theorem 3. *Any lifted tree alignment is an optimal bottleneck tree alignment when the weight function is ultrametric.*

Proof. Let T_o denote an optimal solution of the BTAP. Now, we show any $d(w_i, w_j)$ for $\{w_i, w_j\} \in W$ is less than or equal to the weight of the largest edge of T_o when the weight function is ultrametric. We let v_i be labeled by a sequence w_i and v_j be labeled by a sequence w_j , in which v_i and v_j are the leaves of T . We also let the sequences set $(w_i, w_{i_1}, w_{i_2}, \dots, w_{i_p}, w_j)$ be labeled to the path $P_{i,j} = (v_i, v_{i_1}, v_{i_2}, \dots, v_{i_p}, v_j)$ of T_o . Because the weight function is ultrametric, we have

$$\begin{aligned} d(w_i, w_j) &\leq \max\{d(w_i, w_{i_p}), d(w_{i_p}, w_j)\}, \\ d(w_i, w_{i_p}) &\leq \max\{d(w_i, w_{i_{(p-1)}}), d(w_{i_{(p-1)}}, w_{i_p})\}, \\ &\vdots \\ d(w_i, w_{i_2}) &\leq \max\{d(w_i, w_{i_1}), d(w_{i_1}, w_{i_2})\}. \end{aligned}$$

Hence, $d(w_i, w_j)$ is less than or equal to the weight of largest edge in the path $P_{i,j}$ of T_o . Therefore, we have any $d(w_i, w_j)$ is less than or equal to the weight of the largest edge in T_o . Clearly, the weight of the largest edge of any lifted tree alignment is one of $d(w_i, w_j)$ for all $\{w_i, w_j\} \in W$ and then we can immediately conclude. \square

5 Conclusion

In this paper, we have shown the bottleneck tree alignment problem is NP-complete even when the tree structure is the binary tree and the weight function is metric. Then, we solved this problem in two special cases. It would be interesting to find a constant factor approximation algorithm for the BTAP or whether the BTAP is MAX SNP-hard.

References

1. Aho, A.V.: Algorithms for finding patterns in strings. Handbook of Theoretical Computer Science, Volume A: Algorithms and Complexity **A** (1990), 290–300.
2. Bonizzoni, P. and Vedova, G.D: The complexity of multiple sequence alignment with SP-score that is a metric. Theoretical Computer Science **259** (2001) 63–79.
3. Chan, S.C., Wong, A.K.C and Chiu, D.K.T.: A survey of multiple sequence comparison methods. Bulletin of Mathematical Biology **54** (1992) 563–598.
4. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithm. 2nd edition MIT Press, Cambridge (2001).
5. Gusfield, D.: Algorithms on strings, trees, and sequences: computer science and computational biology. Cambridge University Press Cambridge UK (1997).
6. Higuera, C. and Casacuberta, F.: Topology of strings: median string is NP-complete. Theoretical Computer Science **230** (2000) 39–48.
7. Pevzner, P.A.: Computational Molecular Biology: An Algorithmic Approach. The MIT Press, Cambridge, MA 2000.
8. Ravi, R. and Kececioğlu, J.: Approximation algorithms for multiple sequence alignment under a fixed evolutionary tree. Discrete Applied Mathematics **88** (1998) 355–366.

9. Sankoff, D.: Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics* **28** (1975) 35–42.
10. Sankoff, D. and Kruskal, J.: *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*. Addison-Wesley Reading, MA (1983).
11. Setubal, J. and Meidanis, J.: *Introduction to computational molecular biology*. PWS Publishing Company (1997).
12. Wang, L. and Jiang, J.: On the complexity of multiple sequence alignment. *Journal of Computational Biology* **1** (1994) 337–348.
13. Wang, L., Jiang, T. and Lawler, E.L.: Approximation algorithms for tree alignment with a given phylogeny. *Algorithmica* **16** (1996) 302–315.
14. Wang, L. and Gusfield, D.: Improved approximation algorithms for tree alignment. *Journal of Algorithm* **25** (1997) 255–273.
15. Wang, L., Jiang, J. and Gusfield, D.: A more efficient approximation scheme for tree alignment. *SIAM Journal on Computing* **30** (2000) 283–299.
16. Wagner, R. and Fisher, M.: The string-to-string correction problem. *Journal of the ACM* **21** (1974) 168–178.
17. Wu, B.Y., Chao, K.M. and Tang, C.Y., Approximation and exact algorithms for constructing minimum ultrametric trees from distance matrices. *Journal of Combinatorial Optimization* **3** (1999) 199–211.

Performance Study of a Genetic Algorithm for Sequencing in Mixed Model Non-permutation Flowshops Using Constrained Buffers*

Gerrit Färber and Anna M. Coves Moreno

Instituto de Organización y Control de Sistemas Industriales (IOC),
Universidad Politécnica de Cataluña (UPC), Barcelona, Spain
Gerrit_Faerber@gmx.de,
Anna.Maria.Coves@upc.edu

Abstract. This paper presents the performance study of a Genetic Algorithm applied to a mixed model non-permutation flowshop production line. Resequencing is permitted where stations have access to intermittent or centralized resequencing buffers. The access to the buffers is restricted by the number of available buffer places and the physical size of the products. Characteristics such as the difference between the intermittent and the centralized case, the number of buffer places and the distribution of the buffer places are analyzed. Improvements that come with the introduction of constrained resequencing buffers are highlighted.

1 Introduction

In the classical production line, only products with the same options were processed at once. Products of different models, providing distinct options, were either processed on a different line or major equipment modifications were necessary. For today's production lines this is no longer desirable and more and more rise the necessity of manufacturing a variety of different models on the same line, motivated by offering a larger variety of products to the client. Furthermore, the stock for finished products is reduced considerably with respect to a production with batches, and so are the expenses derived from it.

Mixed model production lines consider more than one model being processed on the same production line in an arbitrary sequence. Nevertheless, the majority of publications in this area are limited to solutions which determine the job sequence before the jobs enter the line and maintain it without interchanging jobs until the end of the production line, which is known as permutation flowshop. Exact approaches for makespan minimization can be found in [1], [2], and [3], among others. In two recent reviews [4], [5] heuristic methods for sequencing problems are presented.

In the case of more than three stations and with the objective function to minimize the makespan, a unique permutation is no longer optimal. In [6] a study of

* This work is partially supported by the Ministry of Science and Technology, and the funding for regional research DPI2004-03472.

the benefit of using non-permutation flowshops is presented. Furthermore, there exist various designs of production lines which permit resequencing of jobs: using large buffers (Automatic-Storage-and-Retrieval-System) which decouple one part of the line from the rest of the line [7]; buffers which are located off-line [8]; hybrid or flexible lines [9]; and more seldomly, the interchange of job attributes instead of physically changing the position of a job within the sequence [10]. Resequencing of jobs on the line is even more relevant with the existence of an additional cost or time, occurring when at a station the succeeding job is of another model, known as setup-cost and setup-time [11].

The present work considers a flowshop with the possibility to resequence jobs between consecutive stations. The buffers are located off-line either accessible from a single station (intermittent case) or from various stations (centralized case). In both cases, it is considered that a job may not be able to be stored in a buffer place, due to its extended physical size, see figure 1.

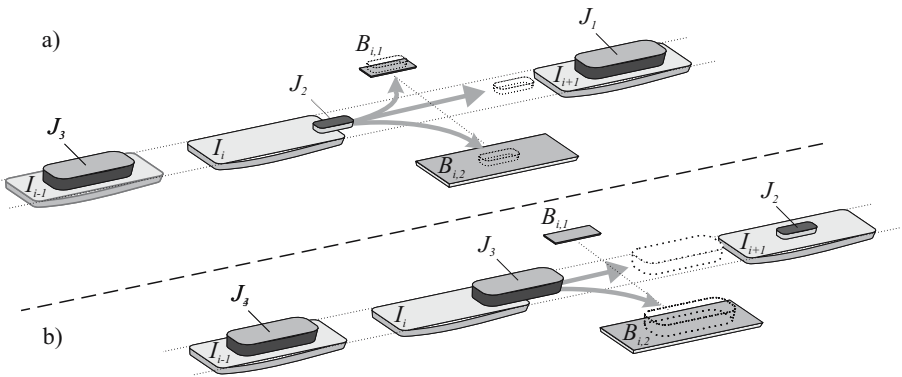


Fig. 1. Scheme of the considered flowshop. The jobs J_j pass consecutively through the stations I_i . The buffer B_i permits to temporally store a job with the objective of reinserting it at a later position in the sequence. a) Job J_2 can pass through any of the two buffer places $B_{i,1}$ or $B_{i,2}$ of buffer B_i . b) Job J_3 can pass only through buffer place $B_{i,2}$, due to its physical size.

The considered problem is relevant to various flowshop applications such as chemical productions dealing with client orders of different volumes and different sized resequencing tanks. Also in productions where split-lots are used for engineering purpose, such as the semiconductor industry. Even in the production of prefabricated houses with, e.g., large and small walls passing through consecutive stations where electrical circuits, sewerage, doors, windows and isolation are applied.

In what follows the problem is formulated with more detail and the applied Genetic Algorithm is described. Thereafter, the accomplished performance study is presented and finally conclusions are presented which are already useful at the time a production line is being designed.

2 Problem Definition

The realized work is based on the classical flowshop in which the jobs ($J_1, J_2, \dots, J_j, \dots, J_n$) pass consecutively through the stations ($I_1, I_2, \dots, I_i, \dots, I_m$). Furthermore, after determined stations, off-line buffers B_i permit to resequence jobs. The buffer provides various buffer places ($B_{i,1}, B_{i,2}, \dots$) and each buffer place is restricted by the physical size of the jobs to be stored. As can be seen in figure 1a, job J_2 can be stored in buffer place $B_{i,1}$ as well as in $B_{i,2}$. Whereas, the next job J_3 can be stored only in buffer place $B_{i,2}$, because of the physical size of the job exceeding the physical size of the buffer place $B_{i,1}$, see figure 1b.

In a first step, the resequencing buffers are located intermittent, between two consecutive stations. In this case the buffer is assigned to the precedent station and may be accessed only by this station. Then, for an additional benefit, a single resequencing buffer is used, with access from various stations, while the limitations on the physical size of the buffer places are maintained.

3 Genetic Algorithm

The concept of the Genetic Algorithm was first formulated by [12] and [13] and can be understood as the application of the principles of evolutionary biology, also known as the survival of the fittest, to computer science. Genetic algorithms are typically implemented as a computer simulation in which a population of chromosomes, each of which represents a solution of the optimization problem, evolves toward better solutions. The evolution starts from an initial population which may be determined randomly. In each generation, the fitness of the whole population is evaluated and multiple individuals are stochastically selected from the current population, based on their fitness and modified to form a new population. The alterations are biologically-derived techniques, commonly achieved by inheritance, mutation and crossover. Multiple genetic algorithms were designed for mixed model assembly lines such as [14], [15], [16] and [17].

The heuristic used here is a variation of the Genetic Algorithm explained in [18]. The genes represent the jobs which are to be sequenced. The chromosomes v , determined by a series of genes, represent a sequence of jobs. A generation is formed by R chromosomes and the total number of generations is G . In the permutation case, the size of a chromosome is determined by the number of jobs, the fraction Π . In the non-permutation case, the chromosomes are $L + 1$ times larger, resulting in the fractions Π_1, \dots, Π_{L+1} , being L the number of resequencing possibilities. In both cases, special attention is required when forming the chromosomes, because of the fact that for each part of the production line every job has to be sequenced exactly one time.

The relevant information for each chromosome is its fitness value (objective function), the number of job changes and the indicator specifying if the chromosome represents a feasible solution. A chromosome is marked unfeasible and is imposed with a penalty, if a job has to be taken off the line and no free buffer place is available or the physical size of the job exceeds the size limitation of the

available buffer places. When two solutions result in the same fitness, the one with fewer job changes is preferred.

3.1 Genetic Operators

The genetic operators specify in which way the subsequent population is generated by reproduction of the present population, taking into account that "fitter" solutions are more promising and therefore are more likely to reproduce. Even an unfeasible solution is able to reproduce, because of the fact that it may generate valuable and feasible solutions in one of the preceding generations. The used genetic operators are inheritance, crossover and mutation. The value p_X is the percentage with which a genetic operator X is applied to a chromosome.

Inheritance: This operator is determined by two parameters. The parameter p_{BS} determines the percentage of the best solutions which will be copied directly to the next generation, called the cluster of promising chromosomes, and ensures that promising chromosomes are not extinct. Then, in order to not remain in a local minimum, the parameter p_b determines the percentage of chromosomes which are removed from this cluster.

Crossover: This operator specifies the operation of interchanging information of two chromosomes. Two crossover operations are applied, crossover-I (figure 2a,b) and crossover-II (figure 2c,d). The probabilities with which these operations are applied to a chromosome are p_{c-I} and p_{c-II} , and the crossover points are defined by the random number pos , and the pair pos_1 and pos_2 , respectively.

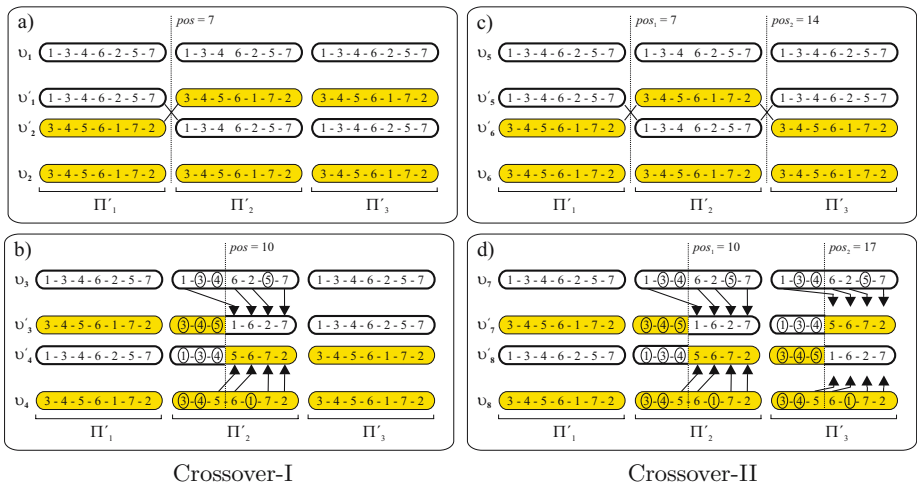


Fig. 2. Operators crossover-I and crossover-II. a) and c) In the simple case the crossing takes place between two main fractions of the chromosome. After the crossover point the chromosomes are completely crossed over. b) and d) In the more complex case it has to be assured that each job is sequenced exactly one time for each fraction of the chromosome.

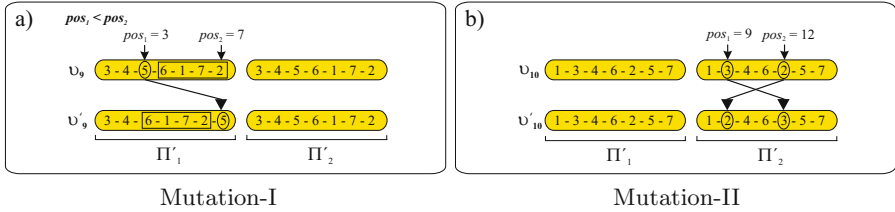


Fig. 3. Operators mutation-I and mutation-II. a) The job at position pos_1 is taken off the line and reinserted to the line at position pos_2 . b) The two jobs at position pos_1 and pos_2 are interchanged.

If the crossover point (pos, pos_1 and pos_2) is a multiple of the number of jobs to be sequenced, the crossover operation is simple and takes place between two main fractions of the chromosome, i.e. after the crossover point the chromosomes are completely crossed over. Whereas, in the complex case the crossover points are located within a main fraction of the chromosome and it has to be assured explicitly that each job is sequenced exactly one time for each fraction of the chromosome.

Mutation: This operator specifies the operation of relocating jobs at position pos_1 to position pos_2 within the same fraction of a chromosome. Two mutation operators are applied, mutation-I and mutation-II (figure 3). Furthermore, there exist two cases for mutation-I: forward mutation, where $pos_1 < pos_2$; and backward mutation, where $pos_1 > pos_2$. In the first case, a single job has to be taken off the line, and in the second case, in order to let a single job pass, a group of succeeding jobs has to be taken off the line, resulting in a larger effort to realize. The probabilities of this operator are $p_{m-I(f)}$, $p_{m-I(b)}$ and p_{m-II} .

3.2 Cascading

In order to further improve the Genetic Algorithm, it is partitioned into two steps. In the first step, the possibility of resequencing jobs within the production line is ignored, furthermore only permutation sequences are considered as possible solutions and the chromosome size is reduced to the number of jobs. The last generation, together with the best found solution, form the initial generation for the next cascade where the resequencing possibilities, provided by stations with access to resequencing buffers, are taken into account.

3.3 Parameter Adjustment

The adjustment of the values of the genetic operators, which are used for the two cascades of the Genetic Algorithm, is determined with an extended experimentation and consists of three steps:

Rough Adjustment: In order to adjust the parameters in a robust manner, different sets of parameters are applied to a series of 14 differently sized problems, varying the number of jobs to be sequenced, the number of stations and

the number of resequencing possibilities. During the rough adjustment only one unique seed is used for the random number generation in the Genetic Algorithm. The sets of parameters are summarized and the 300 most promising which show good performance on all problem sizes are used for further adjustment.

Repeatability: The use of only one seed in the rough adjustment requires to determine amongst the promising sets of parameters which set achieves good results for a multitude of seeds. The fact that a set of parameters achieves good results for different seeds indicates that the same set of parameters also performs well for different plant setups. The promising sets of parameters are verified with 16 different seeds for the 14 differently sized problems. Once the sets of promising parameters are examined with respect to repeatability, one set is used for the fine adjustment, determined by grouping into clusters [19].

Fine Adjustment: Due to the fact that in the previous analysis predetermined discrete values for the parameters are used, a fine adjustment succeeds. The genetic operators are subject to an adjustment of 0.1 for the previously determined sets of parameters and are revised with 16 seeds for the 14 differently sized problems, used for the repeatability.

The experimentation is first performed on the permutation case (first cascade), and then on the non-permutation case (second cascade). Table 1 lists the resulting values of the genetic operators used in the following performance study.

Table 1. Characteristic values of the Genetic Algorithm. The first cascade is applied to determine a generation with only permutation solutions, which is then used as an initial generation for the second cascade.

Cascade	R	G	p_{BS}	p_b	p_{c-I}	p_{c-II}	$p_{m-I(f)}$	$p_{m-I(b)}$	p_{m-II}
Step 1	100	1000	0.05	0.1	0.3	0.6	0.25	0.25	0.25
Step 2	100	10000	0.05	0.4	0.5	0.35	0.45	0.1	0.1

4 Performance Study

For the study of performance, a flowshop which consists of 5 stations is considered. The range of the production time is $[0...20]$ such that for some jobs exists zero-processing time at some stations, for the setup-time $[2...8]$ and for the setup-time $[1...5]$. The number of jobs is varied from 5 to 100 with increments of 5. The objective function, is the weighted sum of the makespan (factor of 1.0) and the setup-cost (factor of 0.3), where the setup-time has is not concerned with a weight but is indirectly included in the calculation of the makespan.

4.1 Intermittent Versus Centralized Location

Replacing the intermittent resequencing buffer places with centralized resequencing buffer places has two benefits. On the one hand, for the case of the same

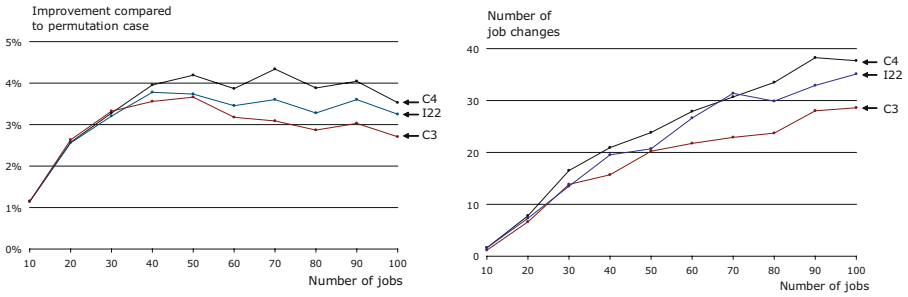


Fig. 4. Comparison of three cases: one centralized resequencing buffer with four buffer places (C4); two intermittent resequencing buffers, each providing two buffer places (I22); one centralized resequencing buffer with three buffer places (C3).

number of buffer places, the objective function of the final solution is expected to be at least as good. This is caused by the fact that in some instances of time, all buffer places of a certain intermittent resequencing buffer may be occupied and do not allow an additional job to be removed from the line, while buffer places from other intermittent resequencing buffers are not accessible. Whereas, in the case of a centralized buffer, blocking only appears when all buffer places are occupied.

On the other hand, the number of buffer places may be reduced in order to obtain values of the objective function similar to the case of the intermittent resequencing buffer. Depending on the number of buffer places which are reduced, this reduction in area is significant.

Figure 4 shows the comparison of the intermittent and the centralized case. After the second station and after the third station there exists access to resequencing buffers. The compared cases are: (I22) two intermittent resequencing buffers, each buffer provides two buffer places; (C3, C4) one centralized resequencing buffer providing three and four buffer places, respectively. On the one hand better solutions are obtained by arranging the buffers centralized; on the other hand, the reduction from four to three buffer places leads to solutions nearly as good as in the intermittent case.

4.2 Number of Buffer Places

The increase in the number of buffer places makes the limitations less strict and as already seen in the previous case, solutions are expected to improve. Figure 5 shows the centralized case without physical size limitations. Jobs leaving the second or the third station have access to the centralized buffer, provided with 2, 3 or 4 buffer places. Providing more buffer places results in better solutions together with an elevated number of job changes.

Figure 6 illustrates an intermittent case. In I22 the second and the third station have 2 buffer places each, in I20 the buffer after the second station has two buffer places and in I02 the buffer after the third station has two buffer places.

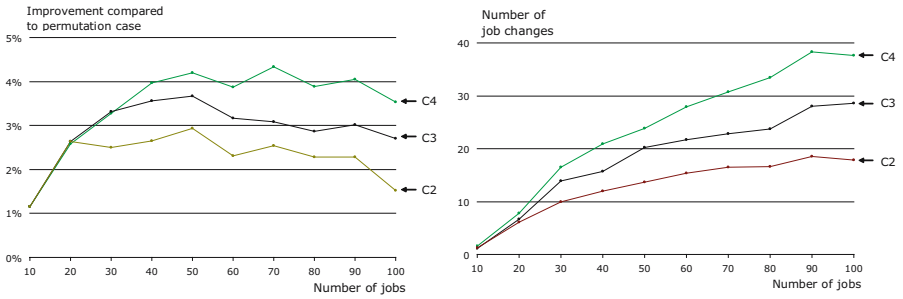


Fig. 5. Variation of the number of buffer places for the centralized case

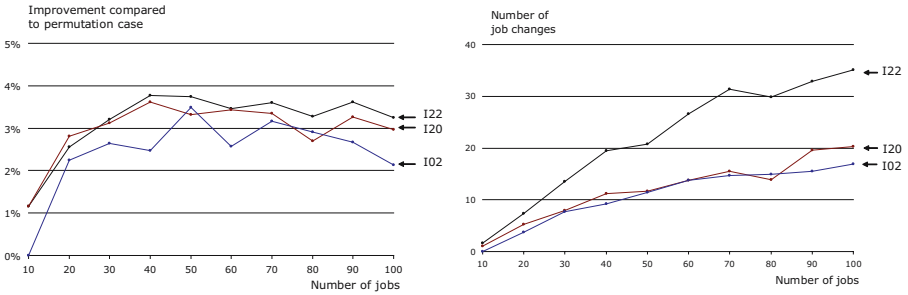


Fig. 6. Variation of buffer places for the case of the intermittent case

Providing two more buffer places results in slightly better solutions together with an elevated number of job changes.

4.3 Difference in Physical Size of Buffer Places

Introducing limitations on the physical size of the buffer places on one side restricts possible solutions but on the other side minimizes the necessary buffer area. This limitation arises, for example, in a chemical production. The arrangement of two tanks which are located off the line, accessible after a certain station, equals an intermittent resequencing buffer with two buffer places. With tank capacities of 50 and 100 liters, a client order of 80 liters can be stored only in the larger of the two tanks which is capable of storing this volume. Whereas, a client order of 50 liters can be stored in either of the tanks. A close look at the local conditions may amortize an increase in the objective function compared to a reduction of investment with respect to tank size and gained area.

As a concrete example, three differently sized buffer places (large, medium, small) are available and the ratio of jobs is $\frac{3}{10}$ large, $\frac{3}{10}$ medium and $\frac{4}{10}$ small. As in the previous section, the second and the third station have access to the resequencing buffers and table 4.3 shows the allocation of the buffer places to the buffers, considering eight scenarios. "300" represents 3 large, 0 medium and 0 small buffer places. In the intermittent case the first buffer is provided

Table 2. Allocation of the buffer places to the buffers. In the intermittent case the allocation is done to two different buffers.

Case	Intermittent			Centralized		
	l	m	s	l	m	s
(300)	1/2	0/0	0/0	3	0	0
(111)	0/1	1/0	0/1	1	1	1
(102)	0/1	0/0	1/1	1	0	2
(012)	0/0	0/1	1/1	0	1	2

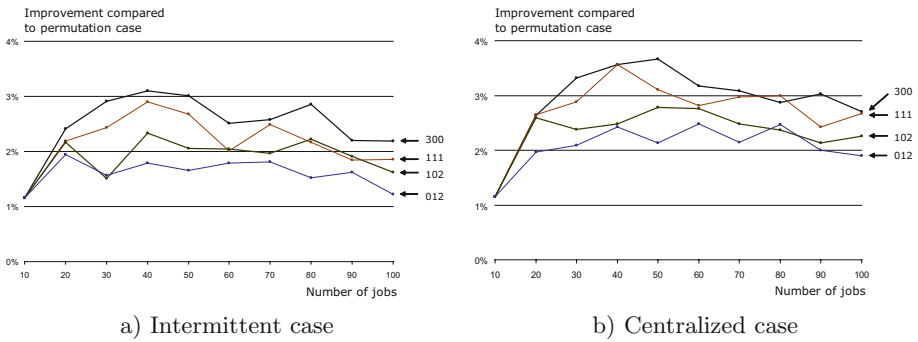


Fig. 7. Influence of the variation of the physical size of the buffer places. "102" represents 1 large, 0 medium and 2 small buffer places. In the intermittent case, the buffer places are divided to two buffers, each with access from a designated station. In the centralized case, the same two stations have simultaneously access to the buffer, containing all three buffer places. The ratio of jobs is $\frac{3}{10}$ large, $\frac{3}{10}$ medium and $\frac{4}{10}$ small.

with 1 and the second buffer with 2 large buffer places. In the centralized case the same two stations have access to a single centralized buffer, containing the three buffer places. Figure 7 shows the influence of the limitation of the physical size. The variation of the size of the buffer places towards smaller buffer places on the one hand decreases the benefit achieved by the possibility of resequencing jobs. On the other hand, it may amortize when taking into account the reduction of investment with respect to tank size and gained area.

5 Conclusions

This paper has presented a study of performance of a genetic algorithm which was applied to a mixed model non-permutation flowshop. The algorithm uses the genetic operators inheritance, crossover and mutation and is designed to consider intermittent or centralized resequencing buffers. Furthermore, the buffer access is restricted by the number of buffer places and the physical size of jobs.

The realized study of performance demonstrates the effectiveness of resequencing by examining certain characteristics. The results of the simulation

experiments reveal the benefits that come with a centralized buffer location, compared to the intermittent buffer location. It either improves the solution or leads to the utilization of fewer resequencing buffer places. An increased number of buffer places clearly improves the objective function and including buffers, constrained by the physical size of jobs to be stored, on one side limits the solutions but on the other side minimizes the necessary buffer area.

In order to take full advantage of the possibilities of resequencing jobs in a mixed model flowshop, additional installations may be necessary to mount, like buffers, but also extra efforts in terms of logistics complexity may arise. The additional effort is reasonable if it pays off the necessary investment. Due to the dependency on local conditions, a general validation is not simple and was not part of this work.

References

- [1] Ignall, E., Schrage, L.: Application of the branch and bound technique to some flow-shop problems. *Operations Research* **13**(3) (1965) 400–412
- [2] Potts, C.: An adaptive branching rule for the permutation flowshop problem. *European Journal of Operational Research* **5**(2) (1980) 19–25
- [3] Carlier, J., Rebai, I.: Two branch and bound algorithms for the permutation flowshop problem. *European Journal of Operational Research* **90**(2) (1996) 238–251
- [4] Framinan, J., Gupta, J., Leisten, R.: A review and classification of heuristics for permutation flowshop scheduling with makespan objective. Technical Report OI/PPC-2001/02 (2002) Version 1.2.
- [5] Framinan, J., Leisten, R.: Comparison of heuristics for flowtime minimisation in permutation flowshops. Technical Report IO-2003/01 (2003) Version 0.5.
- [6] Potts, C., Shmoys, D., Williamson, D.: Permutation vs. non-permutation flow shop schedules. *Operations Research Letters* **10**(5) (1991) 281–284
- [7] Lee, H., Schaefer, S.: Sequencing methods for automated storage and retrieval systems with dedicated storage. *Computers and Industrial Engineering* **32**(2) (1997) 351–362
- [8] Lahmar, M., Ergan, H., Benjaafar, S.: Resequencing and feature assignment on an automated assembly line. *IEEE Transactions on Robotics and Automation* **19**(1) (2003) 89–102
- [9] Engström, T., Jonsson, D., Johansson, B.: Alternatives to line assembly: Some Swedish examples. *International Journal of Industrial Ergonomics* **17**(3) (1996) 235–245
- [10] Rachakonda, P., Nagane, S.: Simulation study of paint batching problem in automobile industry. <http://sweb.uky.edu/~pkrach0/Projects/MFS605Project.pdf> (2000) consulted 14.07.2004.
- [11] Bolat, A.: Sequencing jobs on an automobile assembly line: objectives and procedures. *International Journal of Production Research* **32**(5) (1994) 1219–1236
- [12] Holland, J.: Genetic algorithms and the optimal allocation of trials. *SIAM J. Comput.* **2**(2) (1973) 88–105
- [13] Holland, J.: *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor (1975)

- [14] Bolat, A., Al-Harkan, I., Al-Harbi, B.: Flow-shop scheduling for three serial stations with the last two duplicate. *Computers & Operations Research* **32**(3) (2005) 647–667
- [15] Levitin, G., Rubinovitz, J., Shnits, B.: A genetic algorithm for robotic assembly line balancing. *European Journal of Operational Research* **168** (2006) 811–825
- [16] Wang, L., Zhang, L., Zheng, D.: An effective hybrid genetic algorithm for flow shop scheduling with limited buffers. *Computers & Operations Research* (2006) Article in Press.
- [17] Rubén, R., Concepción, M.: A genetic algorithm for hybrid flowshops with sequence dependent setup times and machine eligibility. *European Journal of Operational Research* **169**(3) (2006) 781–800
- [18] Michalewicz, Z.: *Gentic Algorithms + Data Structures = Evolution Programs*. 3rd edn. Springer Verlag (1996)
- [19] Balasko, B., Abonyi, J., Feil, B.: *Fuzzy clustering and data analysis toolbox; for use with matlab*. (2005)

Optimizing Relative Weights of Alternatives with Fuzzy Comparative Judgment

Chung-Hsing Yeh^{1,2} and Yu-Hern Chang²

¹ Clayton School of Information Technology, Faculty of Information Technology,
Monash University, Clayton, Victoria 3800, Australia
ChungHsing.Yeh@infotech.monash.edu.au

² Department of Transportation and Communications Management,
National Cheng Kung University, Tainan, 701, Taiwan
yhchang@mail.ncku.edu.tw

Abstract. This paper presents an optimal weighting approach for maximizing the overall preference value of decision alternatives based on a given set of weights and performance ratings. In policy analysis settings, relative weights for policy alternatives are subjectively assessed by a group of experts or stakeholders via surveys using comparative judgment. A hierarchical pairwise comparison process is developed to help make comparative judgment among a large number of alternatives with fuzzy ratio values. Performance ratings for policy alternatives are obtained from objective measurement or subjective judgement. The preference value of an expert on a policy alternative is obtained by multiplying the weight of the alternative by its performance rating. Two optimization models are developed to determine the optimal weights that maximize the overall preference value of all experts or stakeholders. An empirical study of evaluating Taiwan's air cargo development strategies is conducted to illustrate the approach.

1 Introduction

Policy analysis and decision making in the public and private sectors often involve the evaluation and ranking of available alternatives or courses of action such as alternative plans, strategies, options or issues. These alternatives are to be evaluated in terms of their weight (relative importance) and performance rating with respect to a single criterion or multiple criteria. In a new decision setting, the weights of the alternatives can only be obtained by expert surveys, while the performance rating can be objectively measured or subjectively assessed by experts.

To obtain the evaluation data by expert surveys, two types of judgment are often required: comparative judgment and absolute judgment [2, 14]. The comparative judgment is required for assessing the relative weight of the alternatives for which a pairwise comparison process is to be used. The absolute judgment can be used to rate the performance of the alternatives independently. To better reflect the inherent subjectiveness and imprecision involved in the survey process, the concept of fuzzy sets [19] is used for representing the assessment results. Modeling using fuzzy numbers has proven to be an effective way for formulating decision problems where the information available is subjective and imprecise [5]. To maximize the overall

preference value assessed by the experts or stakeholders, holding different beliefs and interests about the evaluation problem, we develop an optimal weighting approach.

In subsequent sections, we first present a hierarchical pairwise comparison process for facilitating comparative judgments among a large number of alternatives with fuzzy ratio values. We then develop two optimal weighting models for determining the optimal weights of alternatives for maximizing their overall preference value. Finally we conduct an empirical study to illustrate the optimal weighting approach.

2 Hierarchical Pairwise Comparison for Alternative Weights

To make comparative judgment on the weight (relative importance) of the alternatives, each alternative needs to be compared with all other alternatives. As there are limitations on the amount of information that humans can effectively handle [12], a pairwise comparison approach is advisable to help the experts make comparative judgment. The concept of pairwise comparisons has been popularly implemented in the analytic hierarchy process (AHP) of Saaty [13]. In the AHP, the 1-9 ratio scale is used to compare two alternatives for indicating the strength of their weight. Applying this procedure to all m alternatives will result in a positive $m \times m$ reciprocal matrix with all its elements $x_{ij} = 1/x_{ji}$ ($i = 1, 2, \dots, m; j = 1, 2, \dots, m$).

In this study, we use the 1-9 ratio scale in pairwise comparisons, as it has proven to be an effective measurement scale for reflecting the qualitative information of a decision problem [13]. To reflect the subjectivity and vagueness involved, the ratio value given by the experts is represented by a corresponding triangular fuzzy number. Table 1 illustrates how a triangular fuzzy number is generated to represent the fuzzy assessment from a numeric ratio value assessed by experts. If the ratio value given is 5 (“Strongly more important”), the fuzzy assessment represented as a triangular fuzzy number is (3, 5, 7).

Table 1. Ratio value fuzzification of pairwise comparisons

Equally important	Moderately more important	Strongly more important	Very strongly more important	Extremely more important
1	2	3	4	5
		a_1		a_2
				a_3
				6
				7
				8
				9

In solving a fuzzy positive reciprocal matrix resulting from pairwise comparisons using fuzzy ratios, Buckley [3, 4] uses the geometric mean method to calculate the relative fuzzy values for all the alternatives. This method possesses a number of desirable properties and can be easily extended to fuzzy positive reciprocal matrices. The method guarantees a unique solution to the fuzzy positive reciprocal matrix and can be easily applied to situations where multiple experts are involved in the assessment process [3]. Given a fuzzy positive reciprocal matrix $R = [x_{ij}]$ ($i = 1, 2, \dots, m; j = 1, 2, \dots, m$), the method first computes the geometric mean of each row as

$$r_i = \left(\prod_{j=1}^m x_{ij} \right)^{1/m} \tag{1}$$

The fuzzy values w_i for m alternatives A_i ($i = 1, 2, \dots, m$) are then computed as

$$w_i = r_i / \sum_{j=1}^m r_j \tag{2}$$

With the use of triangular fuzzy numbers, the arithmetic operations on fuzzy numbers are based on interval arithmetic [9]. The Buckley’s method requires the fuzzy positive reciprocal matrix to be reasonably consistent [4, 11].

This pairwise comparison process requires each expert/respondent to make $m(m-1)/2$ comparisons for each attribute. It would be tedious and impractical, if the number of the alternatives to be compared is great. For example, in the empirical study to be presented in Section 4, 153 ($=18(18-1)/2$) comparisons are required for each expert to apply the pairwise comparison process directly to assess the relative importance of all 18 alternative strategies.

To facilitate the pairwise comparison process for assessing a large number of alternatives, we suggest a hierarchical approach, if these alternatives can be grouped into their higher level goals or functional areas. In this case, instead of comparing between all alternatives, the pairwise comparison process is conducted (a) between groups (higher level goals or functional areas), and (b) between alternatives within each group. As such, the number of comparisons required is greatly reduced. For example, when apply this hierarchical pairwise comparison approach to the empirical study, only 36 ($=15+10+1+1+3+3+3$) comparisons are required.

To obtain the relative fuzzy values of all alternatives across all groups, the relative fuzzy values of alternatives under each group are normalized by making the relative fuzzy value of the corresponding group (relative to other groups) as their mean value. If the relative value of a group A_i ($i = 1, 2, \dots, I$) is w_{A_i} and the relative values of its N_i associated alternatives within the group A_i are $w_{A_{ih}}^i$ ($h = 1, 2, \dots, N_i$), then the relative values of these N_i alternatives among all alternatives are

$$w_{A_{ih}} = w_{A_i} \times w_{A_{ih}}^i \times \left(N_i / \sum_{h=1}^{N_i} w_{A_{ih}}^i \right) \tag{3}$$

To defuzzify the fuzzy assessments, we use the concept of α -cut in fuzzy set theory [10]. By using an α -cut on a triangular fuzzy number, a value interval $[x_l^\alpha, x_r^\alpha]$ can be derived, where $0 \leq \alpha \leq 1$. For a given α , x_l^α and x_r^α are the average of the lower bounds and upper bounds of the crisp intervals respectively, resulted from all the α -cuts using the alpha values equal to or greater than the specified value of α . The value of α represents the decision maker’s degree of confidence in the fuzzy assessments of experts/respondents. To reflect the decision maker’s relative preference between x_l^α and x_r^α , an attitude index λ in the range of 0 and 1 can be incorporated. As a result, a crisp value can be obtained as

$$x_\alpha^\lambda = \lambda x_r^\alpha + (1 - \lambda)x_l^\alpha, 0 \leq \lambda \leq 1. \tag{4}$$

The value of λ can be used to reflect the decision maker’s attitude towards risk, which may be optimistic, pessimistic or somewhere in between [16, 17]. In actual decision settings, $\lambda = 1$, $\lambda = 0.5$, or $\lambda = 0$ can be used to indicate that the decision maker has an optimistic, moderate, or pessimistic view respectively on fuzzy assessment results.

3 Optimal Weighting of Alternatives

The weights of alternatives given by the experts or stakeholders may not necessarily be the optimal weights for the evaluation problem as a whole, as these experts may hold different, often conflicting beliefs and interests. In this section, we develop an optimal weighting approach in order to obtain the optimal weights that maximize the overall preference value of all alternatives as a whole. This approach is based on our notion that the interests of the experts or stakeholders on the alternatives are reflected by their preference to these alternatives through their assessments on the weights and performance ratings of alternatives. The preference value of an alternative is obtained by multiplying its weight by its performance rating. If the alternative performance is to be assessed with respect to multiple attributes, a multiattribute decision making method (e.g. [6, 7, 8, 15]) can be used to obtain an overall performance value.

To maximize the total preference value of all alternatives A_i ($i = 1, 2, \dots, m$) for individual experts or stakeholders E_j ($j = 1, 2, \dots, n$), we can use the following model:

Objective

$$\text{Maximize } P_j = \sum_{i=1}^m w_i x_{ij} \tag{5}$$

$$\text{Subject to: } L_i \leq w_i \leq U_i \tag{6}$$

$$\sum_{i=1}^m w_i = 1 \tag{7}$$

$$w_i \geq 0; i = 1, 2, \dots, m; j = 1, 2, \dots, n.$$

- where P_j = the total preference value of expert E_j ($j = 1, 2, \dots, n$);
- w_i = the best weights of alternative A_i ($i = 1, 2, \dots, m$);
- x_{ij} = the performance rating of alternative A_i ($i = 1, 2, \dots, m$) given by expert E_j ($j = 1, 2, \dots, n$);
- L_i = the smallest of weights of alternative A_i ($i = 1, 2, \dots, m$) given by all n experts;
- U_i = the largest of weights of alternative A_i ($i = 1, 2, \dots, m$) given by all n experts.

The objective function (5) is to maximize individual experts’ total preference value, which is represented by multiplying the best importance weights of alternatives (decision variables) by their assessed performance rating. Constraints (6) impose that the best weights generated must lie within the weight ranges assessed by all experts. Constraints (7) state that the best weights obtained for individual experts are to be normalized to sum

to 1. In practical applications, cardinal weights are usually normalized to sum to 1, in order to allow the weight value to be interpreted as the percentage of the total importance weight [1]. Solving Model (5-7) will obtain the best total preference value P_j^* of all alternatives for each individual expert E_j ($j = 1, 2, \dots, n$).

The best weights generated from Model (5-7) reflect the best interests of individual experts or stakeholders, not necessarily the best interests of all experts or stakeholders as a whole. As such, the optimal weights of alternatives should maximize the overall preference value of all experts or stakeholders. In other words, the difference between the best total preference value of individual experts generated by Model (5-7) and the overall preference value generated by the optimal weights should be minimized. This can be achieved by solving the following model:

Objective

$$\text{Minimize } \sum_{j=1}^n (P_j^* - \sum_{i=1}^m W_i x_{ij})^2 \tag{8}$$

$$\text{Subject to: } L_i \leq W_i \leq U_i \tag{9}$$

$$\sum_{i=1}^m W_i = 1 \tag{10}$$

$$W_i \geq 0; i = 1, 2, \dots, m; j = 1, 2, \dots, n.$$

where P_j^* = the best total preference value of expert E_j ($j = 1, 2, \dots, n$) generated by the model (5-7);

W_i = the optimal weights of alternative A_i ($i = 1, 2, \dots, m$);

x_{ij} = the performance rating of alternative A_i ($i = 1, 2, \dots, m$) given by expert E_j ($j = 1, 2, \dots, n$);

L_i = the smallest of weights of alternative A_i ($i = 1, 2, \dots, m$) given by all n experts;

U_i = the largest of weights of alternative A_i ($i = 1, 2, \dots, m$) given by all n experts.

The objective function (8) is to minimize the total preference difference between individual experts' best preference value and the overall preference value of all experts, where W_i ($i = 1, 2, \dots, m$) are the decision variables. Constraints (9) impose that the optimal weights generated for all experts as a whole must lie within the weight ranges assessed by all experts. Constraints (10) state that the optimal weights obtained are to be normalized to sum to 1. Solving Model (8-10) will obtain the optimal weights W_i for all alternative A_i ($i = 1, 2, \dots, m$).

4 Empirical Study

To improve the competitiveness of Taiwan's air cargo industry in the region, and to further develop Taiwan as a regional and international air cargo hub, Taiwan's government has been establishing the national air cargo development policy. Thought workshops and extensive discussions among experts in relevant public and private

sectors, the policy making process has identified 18 alternative development strategies needed for enhancing Taiwan's air cargo industry in order to ensure its sustainable development. Table 2 shows these 18 development strategies, together with their corresponding development goal. With limited resources available, it is of strategic importance for the government to evaluate and prioritize alternative development strategies in terms of their weight (relative importance) and achievability.

Table 2. Development goals and strategies for Taiwan's air cargo industry

Development Goal		Development strategy	
A ₁	Promoting airport competitiveness	A ₁₁	Enhancing airport capacity
		A ₁₂	Applying competitive airport charges
		A ₁₃	Improving operational efficiency
		A ₁₄	Enhancing airport management
		A ₁₅	Enhancing airport transit functions
A ₂	Enhancing integration of transportation systems	A ₂₁	Improving inland transportation and logistics infrastructure
		A ₂₂	Establishing information platform among transportation systems
A ₃	Expanding air cargo routes	A ₃₁	Applying flexible route allocation
		A ₃₂	Lifting carriers' restrictions on the choice of air cargo routes
A ₄	Improving air cargo management systems	A ₄₁	Amplifying air cargo related rules and regulations
		A ₄₂	Strengthening management of hazardous goods
		A ₄₃	Encouraging regional strategic alliance among carriers
A ₅	Developing air cargo operation center	A ₅₁	Expediting the establishment of free trade zones
		A ₅₂	Promoting computerization of air cargo logistics services
		A ₅₃	Fostering internationally qualified professionals
A ₆	Expediting direct air cargo links between Taiwan and China	A ₆₁	Planning facility requirements of direct air cargo services
		A ₆₂	Coordinating technical issues of direct air cargo services
		A ₆₃	Establishing rules and regulations of direct air cargo services

A survey questionnaire was designed to ask the experts in three stakeholder groups (including the government, the general public, and the air cargo industry) to assess (a) the relative weight of 6 development goals and 18 strategies respectively using the

hierarchical pairwise comparison approach presented in Section 2, and (b) the achievability level of these strategies, using a set of self-definable linguistic terms. 37 questionnaire forms were distributed to air cargo experts in Taiwan, and 29 effective responses (including 9 government aviation officials, 11 academic researchers, and 9 air cargo practitioners) were received. It is noteworthy that all these experts are with different organizations, and particularly the 9 practitioners are from different sectors of the air cargo industry. Table 3 shows the survey result of applying Equations (1)-(3), which is the average of the fuzzy assessments given by all the experts.

Table 3. Fuzzy weights and achievability level of development strategies

Development strategy	Relative weight	Achievability level
A_{11}	(0.04, 0.14, 0.50)	(56.5, 66.1, 75.1)
A_{12}	(0.05, 0.17, 0.61)	(60.1, 71.0, 78.3)
A_{13}	(0.08, 0.32, 1.00)	(61.3, 73.7, 80.7)
A_{14}	(0.06, 0.22, 0.75)	(56.9, 68.3, 76.7)
A_{15}	(0.05, 0.18, 0.63)	(59.6, 71.1, 79.7)
A_{21}	(0.03, 0.08, 0.30)	(52.1, 63.2, 72.6)
A_{22}	(0.03, 0.09, 0.33)	(53.4, 64.3, 73.5)
A_{31}	(0.05, 0.18, 0.62)	(49.8, 59.1, 69.0)
A_{32}	(0.04, 0.15, 0.52)	(50.5, 60.7, 70.2)
A_{41}	(0.04, 0.12, 0.43)	(55.0, 65.1, 75.2)
A_{42}	(0.02, 0.05, 0.23)	(50.9, 60.3, 70.2)
A_{43}	(0.04, 0.12, 0.44)	(54.5, 65.9, 74.7)
A_{51}	(0.08, 0.29, 0.95)	(64.6, 75.1, 83.2)
A_{52}	(0.04, 0.14, 0.52)	(59.8, 70.0, 77.9)
A_{53}	(0.05, 0.16, 0.58)	(56.5, 67.6, 75.8)
A_{61}	(0.07, 0.23, 0.73)	(54.8, 66.4, 74.6)
A_{62}	(0.07, 0.24, 0.77)	(55.0, 65.0, 73.8)
A_{63}	(0.08, 0.30, 0.86)	(52.8, 63.3, 72.7)

To ensure the effectiveness and acceptability of the evaluation result, we use Models (5-7) and (8-10) for determining the optimal weights of 18 strategies from the perspective of the air cargo industry in consideration of the views of the government and the general public. As such, the objective function of Model (5-7) is to maximize the interests of the air cargo industry about 18 development strategies, which are represented by the total preference values of 9 practitioner experts (as individual carriers) for these strategies. The preference value of an alternative assessed by an expert in this empirical study is obtained by multiplying the relative weight of the alternative by the achievability level (as the performance rating) of the alternative assessed by the expert. The constraints of Model (5-7) are that the best weights generated must lie within the importance weight ranges assessed by the government and academic experts. Solving the model will obtain the best total preference value P_j^* of all strategies for individual carriers j ($j = 1, 2, \dots, n$).

As a national development policy, the 18 strategies should ideally be prioritized based on the best interests of the air cargo industry as a whole, and not just for any particular carriers. As such, the objective function of Model (8-10) is to maximize the total preference difference between individual carriers' best preference value and the overall preference value of all carriers. The constraints of the model are that the optimal weights generated for all carriers as a whole must lie within the weight ranges assessed by the government and academic experts.

To apply the fuzzy survey results to Models (5-7) and (8-10), we use Equation (4) with $\alpha = 0$ and $\lambda = 0.5$ to reflect that we have no particular preference for the fuzzy assessment results made by the experts. $\alpha = 0$ implies that we use the mean value of a fuzzy number [18]. $\lambda = 0.5$ indicates that we weight all the values derived from fuzzy assessments equally. With the crisp values of relative weights and achievability levels of 18 strategies assessed by 9 government officials, 11 academic researchers, and 9 practitioners (acting as 9 individual carriers), Model (5-7) is solved by using LINDO software systems. The best total preference values (P_j^*) for 9 individual carriers are 54.329, 55.939, 60.654, 69.657, 64.723, 52.944, 77.298, 51.978, and 83.536 respectively. This data are then used to solve Model (8-10) using LINGO software systems. Column 2 of Table 4 shows the optimal weights of 18 strategies. With these optimal weights and the average achievability level accessed by all experts, we can obtain the optimal overall preference value and ranking of 18 strategies as shown in Table 4.

Table 4. Optimal relative weight and preference value of development strategies

Development strategy	Optimal relative weight	Achievability level	Optimal preference value (ranking)
A_{11}	0.051	65.908	3.361 (10)
A_{12}	0.063	69.805	4.398 (8)
A_{13}	0.096	71.885	6.901 (1)
A_{14}	0.078	67.293	5.249 (3)
A_{15}	0.064	70.132	4.488 (6)
A_{21}	0.029	62.615	1.816 (16)
A_{22}	0.027	63.753	1.721 (17)
A_{31}	0.053	59.299	3.143 (11)
A_{32}	0.033	60.477	1.996 (15)
A_{41}	0.043	65.103	2.799 (14)
A_{42}	0.021	60.460	1.270 (18)
A_{43}	0.044	65.017	2.861 (12)
A_{51}	0.083	74.299	6.167 (2)
A_{52}	0.063	69.241	4.362 (7)
A_{53}	0.043	66.655	2.866 (13)
A_{61}	0.058	65.282	3.786 (9)
A_{62}	0.074	64.563	4.778 (5)
A_{63}	0.077	62.914	4.844 (4)

It is noteworthy that the optimal overall preference value of all strategies (i.e. sum of Column 4 of Table 4) is 66.806, which is greater than the average of the best total preference values of all strategies for 9 individual carriers obtained from Model (5-7)

(being 63.451). This suggests that the overall preference value will increase if we consider the best interests of all carriers as a whole, rather than the best interests of individual carriers. This also justifies the use of the optimal weighting approach for the evaluation problem.

In Models (5-7) and (8-10), the relative optimal weights of 18 strategies to be generated for maximizing the best interests of the air cargo industry are constrained by the opinions of the government and academic experts, as given by constraints (6) and (9). In addition to their practical significance in considering all stakeholders' views with focus on the best interests of the air cargo industry, these constraints are necessary for achieving an effective result. If these constraints are not included in the models, the resultant optimal weights of 18 strategies are given in the second column of Table 5. The third and fourth columns of Table 5 show the optimal weights of 18 strategies for the government and academic groups respectively, without considering other stakeholder groups' views. Clearly, the result of the optimal weighting models by considering only the best interests of one stakeholder group is not desirable. It is interesting to note that the three stakeholder groups hold different views of how these strategies should be prioritized in order to maximize their own best interests.

Table 5. Optimal relative weights for three stakeholder groups without constraints

Development strategy	Industry oriented	Government oriented	Academic oriented
A ₁₁	0.000	0.000	0.000
A ₁₂	0.211	0.000	0.000
A ₁₃	0.000	0.318	0.792
A ₁₄	0.000	0.000	0.045
A ₁₅	0.000	0.000	0.000
A ₂₁	0.000	0.000	0.000
A ₂₂	0.000	0.000	0.000
A ₃₁	0.000	0.000	0.000
A ₃₂	0.000	0.000	0.072
A ₄₁	0.000	0.000	0.000
A ₄₂	0.000	0.000	0.000
A ₄₃	0.000	0.000	0.000
A ₅₁	0.789	0.682	0.000
A ₅₂	0.000	0.000	0.000
A ₅₃	0.000	0.000	0.091
A ₆₁	0.000	0.000	0.000
A ₆₂	0.000	0.000	0.000
A ₆₃	0.000	0.000	0.000

5 Conclusion

Evaluating decision alternatives such as courses of action, plans, strategies and policy issues often involves assessing relative weights of decision alternatives via surveys using fuzzy comparative judgment. In this paper we have presented a survey based

optimal weighting approach which can produce optimal weights that maximize the overall preference value of alternatives as a whole. In particular, we have proposed a hierarchical pairwise comparison process for assessing relative weights of a large number of alternatives with comparative judgment. The empirical study conducted has demonstrated the effectiveness of the approach. The approach has general application in optimizing relative weights for assessing policy alternatives based on subjective judgements of different stakeholder groups.

References

1. Belton, V., Stewart, T.J.: *Multiple Criteria Decision Analysis: An Integrated Approach*. Kluwer, Boston (2002)
2. Blumenthal, A.: *The Process of Cognition*. Prentice-Hall, Englewood Cliffs, New Jersey (1977)
3. Buckley, J.J.: Fuzzy Hierarchical Analysis. *Fuzzy Sets and Systems* **17** (1985) 233-247
4. Buckley, J.J., Feuring, T., Hayashi, Y.: Fuzzy Hierarchical Analysis Revisited. *European Journal of Operational Research* **129** (2001) 48-64
5. Chang, Y.-H., Yeh, C.-H.: A Survey Analysis of Service Quality for Domestic Airlines. *European Journal of Operational Research* **139** (2002) 166-177
6. Chang, Y.-H., Yeh, C.-H.: A New Airline Safety Index. *Transportation Research Part B: Methodological* **38** (2004) 369-383
7. Dyer, J.S., Fishburn, P.C., Steuer, R.E., Wallenius, J., Zionts, S.: Multiple Criteria Decision Making, Multiattribute Utility Theory: The Next Ten Years. *Management Science* **38** (1992) 645-653
8. Hwang, C.L., Yoon, K.: *Multiple Attribute Decision Making - Methods and Applications*. Springer-Verlag, New York (1981)
9. Kaufmann, A., Gupta, M.M.: *Introduction to Fuzzy Arithmetic, Theory and Applications*. International Thomson Computer Press, Boston (1991)
10. Klir, G.R., Yuan, B.: *Fuzzy Sets and Fuzzy Logic Theory and Applications*. Prentice-Hall, Upper Saddle River, NJ (1995)
11. Leung, L.C., Cao, D.: On Consistency and Ranking of Alternatives in Fuzzy AHP. *European Journal of Operational Research* **124** (2000) 102-113
12. Miller, G. A.: The Magic Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review* **63** (1956) 81-97
13. Saaty, T.L.: *The Analytic Hierarchy Process*. McGraw-Hill, New York (1980)
14. Saaty, T.L.: Rank from Comparisons and from Ratings in the Analytic Hierarchy/Network Processes. *European Journal of Operational Research* **168** (2006) 557-570
15. Yeh, C.-H.: The Selection of Multiattribute Decision Making Methods for Scholarship Student Selection. *International Journal of Selection and Assessment* **11** (2003) 289-296
16. Yeh, C.-H., Deng, H., Chang, Y.-H.: Fuzzy Multicriteria Analysis for Performance Evaluation of Bus Companies. *European Journal of Operational Research* **26** (2000) 1-15
17. Yeh, C.-H., Kuo, Y.-L.: Evaluating Passenger Services of Asia-Pacific International Airports. *Transportation Research Part E* **39** (2003) 35-48
18. Yeh, C.-H., Deng, H.: A Practical Approach to Fuzzy Utilities Comparison in Fuzzy Multi-Criteria Analysis. *International Journal of Approximate Reasoning* **35** (2004) 179-194
19. Zadeh, L.A.: Fuzzy Sets. *Information and Control* **8** (1965) 338-353

Model and Solution for the Multilevel Production-Inventory System Before Ironmaking in Shanghai Baoshan Iron and Steel Complex

Guoli Liu and Lixin Tang*

The Logistics Institute, Northeastern University, 110004 Shenyang, China
Tel.: +86-24-83680169; Fax: +86-24-83680169
qhjytlx@mail.neu.edu.cn, liuguoliguohui@yahoo.com.cn

Abstract. This research deals with the production-inventory problem originating from the ironmaking production system in Shanghai Baoshan Iron and Steel Complex (Baosteel). To solve this multilevel, multi-item, multi-period, capacitated lot-sizing problem, a deterministic mixed integer programming (MIP) model based on the minimization of production and inventory costs is formulated to determine the production and inventory quantities of all materials in each time period under material-balance and capacity constraints. Due to the large-scale variables and constraints in this model, a decomposition Lagrangian relaxation (LR) approach is developed. Illustrative numerical examples based upon the actual production data from Baosteel are given to demonstrate the effectiveness of the proposed LR approach.

Keywords: Lot-sizing, combinatorial optimization, lagrangian relaxation.

1 Introduction

Effective multilevel production-inventory planning has become the focus of research efforts. A wide variety of models and various solution procedures have been proposed. Diaby *et al.* [1] developed a Lagrangean relaxation-based heuristic procedure to generate near-optimal solutions to very-large-scale capacitated lot-sizing problems (CLSP) with setup times and limited overtime. Afentakis and Gavish [2] developed algorithms for optimal lot sizing of products with a complex product structure. They converted the classical formulation of the general structure problem into a simple but expanded assembly structure with additional constraints, and solved the transformed problem by a branch-and-bound based procedure. Afentakis *et al.* [3] presented a new formulation of the lot-sizing problem in multistage assembly systems, which led to an effective optimization algorithm for the problem. Billington *et al.* [4] dealt with the capacitated, multi-item, multi-stage lotsizing problem for serial systems by examining the performance of heuristics found effective for the capacitated multiple-product, single-stage problem. Billington *et al.* [5] introduced a line of research on capacity-constrained multi-stage production scheduling problems. A review of the literature and an analysis of the type of problems that existed were presented. In their

* Corresponding author.

subsequent paper (Billington *et al.* [6]) they presented a heuristic method, based on Lagrangian relaxation, for multilevel lot-sizing when there was a single bottleneck facility.

This paper investigates the problem of making the production and inventory decisions for ironmaking materials manufactured through several stages in Baosteel so as to minimize the sum of setup, production and inventory holding costs while the given demand in each time period is fulfilled without backlogging. After an investigation on the multilevel production-inventory system before ironmaking in Baosteel has been made, the authors of this paper conclude that the process of making hot metals for the steel plant should be viewed as a continuous line process of consecutive batch operations that are illustrated in Figure 1. The problem considered here is virtually a deterministic, dynamic-demand, multilevel, multi-item, multi-period, multi-pass, capacitated lot-sizing problem in a special system structure depicted as a cyclic graph, which could be detected from the features summarized in Table 1.

The remainder of this paper is organized as follows. In section 2, a novel mixed integer programming model is formulated to solve the production-inventory problem originating from Baosteel. In section 3, the Lagrangian relaxation approach is presented and the computational results are reported. Finally, Section 4 presents a discussion of future work and concludes the paper.

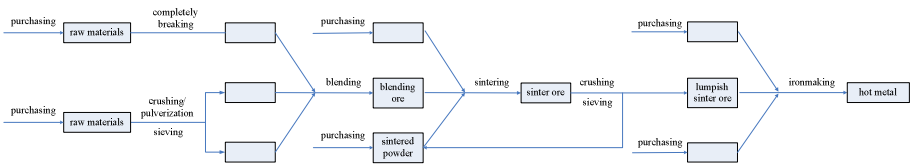


Fig. 1. The flowsheet of the ironmaking production system in Baosteel

Table 1. Features of the production-inventory problem in Baosteel

-
- Each material may have several predecessors and several successors.
 - The capacity constraints may include the capacity limitations of multiple resources.
 - Independent demands could only be given to the final products.
 - Each material could respond to more than one technology routing.
 - The structure of the production system is allowed to be cyclic.
 - Both assembly and disassembly structures exist in the production system.
 - No backlogging is allowed.
-

2 Preliminaries and Mathematical Formulation of the Problem

For the production-inventory problem considered in this paper, the following assumptions apply:

- (1) The initial inventories of all items are zero.
- (2) No backorder allowed.
- (3) One material or material group could be changed into multiple products on a one-to-many or many-many operation.
- (4) One material or material group could be used to produce one product on a one-one or many-to-one operation.
- (5) In each time period, every intermediate product could be produced on different operations.
- (6) On one operation each intermediate product can only be produced from entirely different materials or material groups in any time period.
- (7) Some by-products could be processed to reversely supply their predecessors.
- (8) Each final product represents one kind of hot metal;
- (9) The external demands of final products are given and change with time.
- (10) No inventory problems on the final products are considered.
- (11) Each operation used to directly produce final products is denoted as '0' and the operation of purchasing as '1'.
- (12) Operation "1" is considered as a one-one operation.
- (13) Every operation denoted as "0" should be a many-to-one or one-one operation.
- (14) Each final product only has independent demands.
- (15) Only one sequence of operations is appointed to produce every final product, so there is no alternative.

Parameters in the model are:

K = the number of stock yards;

T = the set of indices for time periods;

N = the set of indices for items in the production process, excluding final products;

N_f = the set of final products;

L = the set of indices for all operations, excluding reversely supplying;

L_1 = the set of one-to-many and many-many operations in L ;

L_2 = the set of one-one and many-to-one operations in L , i.e., $L_2 = L - L_1$;

L_R = the operation standing for reversely supplying;

F_{lt} = the set of the core products of operation l in period t , $l \in L_1$, $t \in T$;

B_{ilt} = the set of the by-products corresponding to core product i of operation l in period t , $l \in L_1$, $t \in T$, $i \in F_{lt}$;

J_{lt} = the set of the products of operation l in period t , $l \in L_2$, $t \in T$;

D_{it} = the demand for final product i in period t , $i \in N_f$, $t \in T$;

S_i = the set of immediate successors of item i , $i \in N$;

h_i = the holding cost for one unit of item i in a period, $i \in N$;

$pr_{it} = \{l \in L \mid i \in J_{lt} \text{ or } i \in F_{lt} \text{ or } i \in B_{ilt}\}$, $i \in N$, $t \in T$;

A_{lt} = the set of items needed to be processed on operation l in period t , $l \in L \cup \{L_R\}$, $t \in T$;

$reverse_i$ = the set of items which could be provided by item i on operation L_R , $i \in N$;

$$H_{ijt} = \{l \in pr_{jt} \mid i \in A_{lt}\} \cup \{L_R \mid j \in reverse_i\}, i \in N, j \in S_i, t \in T;$$

$back_i$ = the set of items which could provide item i on operation L_R , $i \in N$;

$lead_{ij}$ = the lead time of item j provided by item i on operation L_R , $i \in N, j \in reverse_i$;

ρ_{ij} = the total quantity of item i used to produce one unit of item j , $i \in N, j \in reverse_i$;

R_{ilt} = the setup cost for the production of item i on operation l in period t , $i \in N, l \in L, t \in T$;

ms_{ilt} = the unit production cost for item i on operation l in period t , $i \in N, l \in L, t \in T$;

r_{ijlt} = the total quantity of item i used to produce one unit of item j on operation l in period t , $l \in L_1, t \in T, i \in A_{lt}, j \in F_{lt} \cap S_i$;

$r_{ijlt} = 0, l \in L_1, t \in T, i \in A_{lt}, j \in B_{lt} \cap S_i$;

$r_{ij0t} = 0, l \in L_1, t \in T, i \in A_{lt}, j \in B_{lt} \cap S_i$;

r_{ijlt} = the total quantity of item i used to produce one unit of item j on operation l in period t , $l \in L_2, t \in T, i \in A_{lt}, j \in S_i$;

r_{ijlt} = the total quantity of item i simultaneously obtained when one unit of item j is produced on operation l in period t , $l \in L_1, t \in T, i \in F_{lt}, j \in B_{lt}$;

p_{ilt} = the capacity absorption coefficient of item i in period t for operation l , $l \in L_2, t \in T, i \in J_{lt}$;

p_{ilt} = the total quantity of the immediate predecessor needed to produce one unit of item i on operation l in period t , $l \in L_1, t \in T, i \in F_{lt}$;

P_{lt} = the capacity limit for operation l in period t , $l \in L \setminus \{1\}, t \in T$;

P_{1t} = the purchasing budget in period t , $t \in T$;

V_k = the largest inventory capacity of stock yard k , $1 \leq k \leq K$;

M = a very large positive number.

Decision variables in the model are:

x_{it} = the production quantity for item i in period t , $i \in N, t \in T$;

I_{it} = the inventory position for item i at the end of period t , $i \in N, t \in T$;

z_{ilt} = the production quantity for item i on operation l in period t , $i \in N, l \in L, t \in T$;

y_{ilt} = the binary variable which takes the value 1 if a setup is made for item i on operation l in period t and zero otherwise, $i \in N, l \in L, t \in T$;

w_{ijt} = the total quantity of item j which is provided by item i on operation L_R and can be used in period t , $i \in N, j \in reverse_i, t \in T$.

Using the above definitions, the production-inventory problem can now be stated as:

Minimize

$$C \equiv \sum_{t \in T} \sum_{l \in L_1 \cap pr_{it}} \sum_{i \in F_{lt}} (R_{ilt} \cdot y_{ilt} + ms_{ilt} \cdot z_{ilt}) + \sum_{t \in T} \sum_{l \in L_2 \cap pr_{it}} \sum_{i \in J_{lt}} (R_{ilt} \cdot y_{ilt} + ms_{ilt} \cdot z_{ilt}) + \sum_{i \in N} \sum_{t \in T} \frac{1}{2} \cdot h_i \cdot (I_{i,(t-1)} + x_{it} + I_{it}) \tag{1}$$

Subject to

$$I_{i,(t-1)} + x_{it} - I_{it} - \sum_{j \in S_i} \sum_{l \in H_{ij} \setminus \{L_R, 0\}} r_{ijl} \cdot z_{jlt} \tag{2}$$

$$= \sum_{j \in S_i \cap N_f} r_{ij0t} \cdot D_{jt} + \sum_{j \in reverse_i} \rho_{ij} \cdot w_{i,j,t+lead_{ij}}, \forall i \in N, t \in T$$

$$z_{klt} = r_{iklt} z_{ilt}, \forall k \in B_{ilt}, i \in F_{lt}, l \in L_1, t \in T \tag{3}$$

$$z_{ilt} \leq M \cdot y_{ilt}, \forall i \in N, l \in pr_{it}, t \in T \tag{4}$$

$$\sum_{i \in F_l} p_{ilt} \cdot z_{ilt} \leq P_l, \forall l \in L_1, t \in T \tag{5}$$

$$\sum_{i \in J_h} p_{ilt} \cdot z_{ilt} \leq P_h, \forall l \in L_2, t \in T \tag{6}$$

$$\sum_{l \in pr_i} z_{ilt} + \sum_{j \in Back_i} w_{j,i,t} = x_{it}, \forall i \in \{i | pr_i \neq \emptyset \text{ or } Back_i \neq \emptyset, i \in N\}, t \in T \tag{7}$$

$$\frac{1}{2} \cdot \sum_{i \in U_k} (I_{i,(t-1)} + x_{it} + I_{it}) \leq V_k, k = 1, \dots, K, \forall t \in T \tag{8}$$

$$I_{i0} = 0, \forall i \in N \tag{9}$$

$$I_{it}, x_{it}, z_{ilt} \geq 0, \forall i \in N, l \in pr_{it}, t \in T \tag{10}$$

$$y_{ilt} \in \{0, 1\}, \forall i \in N, l \in pr_{it}, t \in T \tag{11}$$

The objective function in (1) represents the total setup, production/purchasing and inventory holding cost for all items over all time periods. Constraints (2) are material balance equations which state that dependent and independent demands are fulfilled from inventory, production and reverse logistics. Constraints (3) are defined to ensure that each core product and its by-products are produced in proportion. Constraints (4) together with Constraints (11) indicate that a setup cost will be incurred when a lot size is produced. Constraints (5) and (6) enforce that production must be within the limitations set by available capacity. Constraints (7) imply that each material is supplied by multiple production routings including reverse logistics. Constraints (8) represent inventory capacity limits. Constraints (9) establish zero initial inventories. Constraints (10) define the value ranges of the variables.

3 Solution Methodology

A solution strategy composed of Lagrangian relaxation, linear programming and heuristics is developed. The details for the strategy are presented below.

3.1 Lagrangian Relaxation Algorithm

The decomposition approach for the production-inventory problem considered in this paper hinges on the observation that when we disregard constraints (4), the Lagrangian relaxation problem can be decomposed into easily solvable sub-problems for each kind of variables. With these facts, constraints (4) are dualized in the objective function with non-negative multipliers $\{u_{ilt}\}$:

(LR)

Minimize

$$C_{LR}(u_{ilt}) \equiv \sum_{i \in T} \sum_{l \in L_1 \cap pr_{it}} \sum_{i \in F_{it}} (R_{ilt} \cdot y_{ilt} + mS_{ilt} \cdot z_{ilt}) + \sum_{i \in T} \sum_{l \in L_2 \cap pr_{it}} \sum_{i \in J_{it}} (R_{ilt} \cdot y_{ilt} + mS_{ilt} \cdot z_{ilt}) + \sum_{i \in N} \sum_{t \in T} \frac{1}{2} \cdot h_i \cdot (I_{i,(t-1)} + x_{it} + I_{it}) + \sum_{i \in N} \sum_{t \in T} \sum_{l \in pr_{it}} u_{ilt} \cdot (z_{ilt} - M \cdot y_{ilt}) \tag{12}$$

Subject to constraints (2), (3), (5)-(11) and

$$u_{ilt} \geq 0, \forall i \in N, l \in pr_{it}, t \in T \tag{13}$$

The relaxed problem (LR) can be decomposed into two smaller sub-problems, each involving one type of variables. The sub-problems are given as follows.

(LR₁)

Minimize

$$C_{LR_1}(u_{ilt}) \equiv \sum_{i \in T} \sum_{l \in L_1 \cap pr_{it}} \sum_{i \in F_{it}} R_{ilt} \cdot y_{ilt} + \sum_{i \in T} \sum_{l \in L_2 \cap pr_{it}} \sum_{i \in J_{it}} R_{ilt} \cdot y_{ilt} - M \cdot \sum_{i \in N} \sum_{t \in T} \sum_{l \in pr_{it}} u_{ilt} \cdot y_{ilt} \tag{14}$$

Subject to constraints (11) and (13)

(LR₂)

Minimize

$$C_{LR_2}(u_{ilt}) \equiv \sum_{i \in T} \sum_{l \in L_1 \cap pr_{it}} \sum_{i \in F_{it}} mS_{ilt} \cdot z_{ilt} + \sum_{i \in T} \sum_{l \in L_2 \cap pr_{it}} \sum_{i \in J_{it}} mS_{ilt} \cdot z_{ilt} + \sum_{i \in N} \sum_{t \in T} \sum_{l \in pr_{it}} u_{ilt} \cdot z_{ilt} + \sum_{i \in N} \sum_{t \in T} \frac{1}{2} \cdot h_i \cdot (I_{i,(t-1)} + x_{it} + I_{it}) \tag{15}$$

Subject to constraints (2), (3), (5)-(10), (13)

Solution of the following Lagrangian Dual problem gives the maximum lower bound.

(LD)

Maximize $C_D(u_{ilt}) \equiv \min C_{LR}$ (16)

Subject to constraints (2), (3), (5)-(11) and (13)

3.2 Sub-problem Solution

(LR₁)

Minimize

$$\begin{aligned}
 C_{LR_1}(u_{ilt}) &\equiv \sum_{t \in T} \sum_{l \in L_1 \cap pr_{it}} \sum_{i \in F_{it}} R_{ilt} \cdot y_{ilt} + \sum_{t \in T} \sum_{l \in L_2 \cap pr_{it}} \sum_{i \in J_{it}} R_{ilt} \cdot y_{ilt} - M \cdot \sum_{i \in N} \sum_{t \in T} \sum_{l \in pr_{it}} u_{ilt} \cdot y_{ilt} \\
 &= \sum_{t \in T} \sum_{l \in L_1 \cap pr_{it}} \sum_{i \in F_{it}} (R_{ilt} - M \cdot u_{ilt}) \cdot y_{ilt} + \sum_{t \in T} \sum_{l \in L_2 \cap pr_{it}} \sum_{i \in J_{it}} (R_{ilt} - M \cdot u_{ilt}) \cdot y_{ilt} \\
 &\quad - M \cdot \sum_{t \in T} \sum_{l \in L_1 \cap pr_{it}} \sum_{i \in N \setminus F_{it}} u_{ilt} \cdot y_{ilt} - M \cdot \sum_{t \in T} \sum_{l \in L_2 \cap pr_{it}} \sum_{i \in N \setminus J_{it}} u_{ilt} \cdot y_{ilt} \tag{17}
 \end{aligned}$$

Subject to constraints (11) and (13)

Recall that y_{ilt} is a zero-one variable. To minimize the objective function described by (17), y_{ilt} equals zero when its coefficient is positive and one otherwise. Since LR₂ is a typical linear programming model, it can be directly solved by standard linear programming software package OSL.

3.3 Construction of a Feasible Solution to the Original Problem

Because the discrete decision variables and the continuous decision variables are calculated separately in different sub-problems, the consistency among them defined by constraints (4) is generally violated. To construct a feasible solution based on the optimal solution of LR₂, a heuristic method is adopted to restore feasibility. The algorithmic steps of the heuristics are outlined as follows.

- Step 1.* Take the optimal solution of LR₂ as the initial solution.
- Step 2.* Search for item i that meet the conditions: $i \in I$ and $visit_mark [i] = 0$. If there exist no such items, go to *Step 5*.
- Step 3.* Search for the operation that satisfies the condition: $z_{ilt} > 0$. Set $y_{ilt} = 1$.
- Step 4.* If there exists no such operation, set $visit_mark [i] = 1$, go to *Step 2*; otherwise go to *Step 3*.
- Step 5.* Stop. A feasible solution $\{ I_{it}, x_{it}, z_{ilt}, y_{ilt} \}$ is obtained.

3.4 Subgradient Algorithm for Dual Maximization

In order to solve the Lagrangian dual problem with respect to (16), a subgradient method is constructed in what follows. The vector of multipliers, u_{ilt} , is updated by

$$u_{ilt}^{m+1} = \text{Max}\{0, u_{ilt}^m + t^m \theta^m(u_{ilt}^m)\} \tag{18}$$

where t^m is the step size at the m -th iteration and $\theta^m(u_{ilt}^m)$ is the subgradient of $C_{LR}(u_{ilt})$. The subgradient component relating the i -th material, the l -th operation and the t -th time period is $\theta_{ilt}^m(u_{ilt}^m) = z_{ilt} - M \cdot y_{ilt}$. The step size t^m is given by

$$t^m = \beta_m \frac{C^U - C^m}{\|\theta^m\|^2}, \quad 0 < \beta_m < 2 \tag{19}$$

where C^U is the best objective value found so far and C^m is the value of $C_{LR}(u_{it})$ at the m -th iteration. The parameter β_m is initially set to be a value greater than 0 and is multiplied by a factor after each iteration. The algorithm terminates when a given iteration number has been executed or a fixed duality gap has been reached.

3.5 Computational Results

The algorithms described in previous sections were programmed in Visual C++. In this experiment, the ironmaking production system in Baosteel described in Figure 1 was considered. To generate representative problem instances, the problem parameters in this experiment are based on actual production data from the ironmaking production system in Baosteel. Since the maximum number of materials used in practice is always less than 75, the performance of those algorithms was evaluated on a set of 120 test problems ranging in size from 15 to 75 materials and 6 to 18 time periods. Because Lagrangian relaxation cannot guarantee optimal solutions, the relative dual gap $(C^{UB} - C^{LB})/C^{LB}$ was used as a measure of solution optimality, where C^{UB} is the upper bound to the original problem and C^{LB} is the lower bound. Optimal performance and running times of the Lagrangian relaxation method against different problem sizes are presented in Table 2. For all the algorithms tested, the same iteration number, 100, is imposed on the stopping criterion.

Table 2. Average duality gaps and the average running times

Problem No.	Problem structures Materials × Periods	Average Duality gap (%)	Running time (s)
1	15 × 6	7.747066	1.552000
2	35 × 6	8.826898	6.668000
3	55 × 6	8.910097	40.339999
4	75 × 6	9.202016	28.589001
5	15 × 12	8.143565	4.872000
6	35 × 12	7.528647	22.871001
7	55 × 12	9.590387	53.315997
8	75 × 12	9.129756	93.825000
9	15 × 18	9.229039	9.835000
10	35 × 18	7.022790	49.191000
11	55 × 18	7.515264	116.329004
12	75 × 18	7.639370	217.553003
Average		8.373741	53.74508

From the computational results, the following observations can be made.

- (1) The average duality gaps of all algorithms tested are below 10% in all cases.
- (2) The computational time increases as the number of materials increases.
- (3) The computational time increases as the number of periods increases.
- (4) The duality gaps are relatively stable when the number of materials or periods increases.

4 Conclusions

In this paper, we discussed the production-inventory problem derived from the iron-making production system in Baosteel and formulated it as a separable mixed integer programming model in order to determine the production and inventory quantity of each material in each time period under complicated material-balance constraints and capacity constraints. The objective of the model was to minimize the total setup, production/purchasing and inventory holding cost over the planning horizon. A decomposition solution methodology composed of Lagrangian relaxation, linear programming and heuristics was applied to solve the problem. Computational results for test problems generated based on the production data of Baosteel indicated that the proposed Lagrangian relaxation algorithm can always find good solutions within a reasonable time. Our further research may focus on extending the MIP model to other industries.

Acknowledgement

This research is partly supported by National Natural Science Foundation for Distinguished Young Scholars of China (Grant No. 70425003), National Natural Science Foundation of China (Grant No. 60274049) and (Grant No. 70171030), Fok Ying Tung Education Foundation and the Excellent Young Faculty Program of the Ministry of Education, China. The authors would like to thank the Production Manufacturing Center in Baosteel for providing a lot of production information and data.

References

1. Diaby, M., Bahl, H.C., Karwan, M.H., Zionts, S.: A Lagrangean relaxation approach for very-large-scale capacitated lot-sizing. *Management Science*. 38(9) (1992) 1329-1340
2. Afentakis, P. and Gavish, B.: Optimal lot-sizing algorithms for complex product structures. *Operations Research*. 34(2) (1986) 237-249
3. Afentakis, P., Gavish, B., Karmarkar, U.: Computationally efficient optimal solutions to the lot-sizing problem in multistage assembly systems. *Management Science*. 30(2) (1984) 222-239
4. Billington, P., Blackburn, J., Maes, J.: R. Millen and L.N.V. Wassenhove, Multi-item lot-sizing in capacitated multi-stage serial systems. *IIE Transactions*. 26(2) (1994) 12-24
5. Billington, P.J., McClain, J.O., Thomas, L.J.: Mathematical programming approaches to capacity-constrained MPR systems: Review, Formulation and Problem reduction. *Management Science*. 29(10) (1983) 1126-1141
6. Billington, P.J., McClain, J.O., Thomas, L.J.: Heuristics for multilevel lot-sizing with a bottleneck. *Management Science*. 32(8) (1986) 989-1006

A Coordination Algorithm for Deciding Order-Up-To Level of a Serial Supply Chain in an Uncertain Environment

Kung-Jeng Wang¹, Wen-Hai Chih¹, and Ken Hwang²

¹Department of Business Administration, National Dong Hwa University,
Hualien, 974, Taiwan, R.O.C.

kjwang@mail.ndhu.edu.tw

²Graduate Institute of Technology and Innovation Management,
National Chengchi University, Taipei, Taiwan, R.O.C.

Abstract. This study proposed a method for coordinating the order-up-to level inventory decisions of isolated stations to cooperatively compensate loss in supply chain performance in a serial supply chain due to unreliable raw material supply. Doing so will improve the customer refill rate and lower the production cost within an uncertain environment in which the chain is highly decentralized, both material supply and customer demand are unreliable, and delay incurs in information flow and material flow. The proposed coordination method is then compared to single-station and multiple-station order-up-to policies. Simulated experiments reveal that our proposed method outperforms the others and performs robustly within a variety of demand and supply situations.

1 Introduction

A supply chain is regarded as a large and highly distributed system in the study. Each member in the chain is geographically situated and adheres to distinct management authorities. Centralized control can deteriorate responsiveness and, thus, is inapplicable to such a chain system. A coordinated policy is required to combat fluctuating supply and demand as well as information and material delays within a distributed supply chain. This, in turn, can improve customer satisfaction and reduce supply chain operating costs. Commonly, members in a decentralized supply chain employ installation stock policy to determine ordering quantity. However, such decision-making is myopic and has no optimization guarantee. This motivates many researchers to pursue more efficient way to deal with the issue of order-up-to level decision-making in a decentralized manner.

A serial type of supply chain is the focus of this study. A serial supply chain is the essence of a generalized supply chain network. Research regarding managing a non-fully cooperative, distributed serial chain and the serial supply chain inventory optimization with multi-criteria and asymmetric information remains very active [1,2].

This study proposed a novel order-up-to level (OUTL) inventory policy to against demand and material supply uncertainty and incomplete information situations. Typically, a supply chain is quantitatively evaluated from the perspectives of customer responsiveness (refill rates) and cost [3]. Herein, a synthesis index of these two

measures is applied. Furthermore, the proposed policy is compared to two heuristics, single-station and multiple-station policies [4]. Fuzzy membership function is employed to represent material supply and customer demand within an uncertain environment. Simulation experiments were conducted to examine the robustness of this policy within a variety of supply and demand changes.

The rest of this paper is organized as follows. Literature survey is presented in section 2. Section 3 defines the necessary notations. In section 4, we present a coordinated order-up-to level inventory policy and its computational algorithm. We illustrate the algorithm using a serial chain example with five stations. Section 5 evaluates the performance of the proposed coordinated OUTL policy, using a simulation model. The paper summaries with a concluding remark.

2 Literature Review

Fundamentally, a supply chain consists of procurement, production, and distribution functionalities. Within each of these units, some degree of uncertainty exists. For instance, material shortage and defect, and damage during transportation can disturb procurement, stock shortage, lead time and lot sizing delays can interrupt production, and demand fluctuations can harm distribution tasks. Fuzzy membership function has been used to represent material supply and customer demand within an uncertain environment [4,5]. To manage inventory control within a serial supply chain, [6] proposed two order-up-to policies, single-station and multiple-station approaches within a distributed information environment. To reduce global supply chain cost and to increase customer satisfaction, inventory control systems within a supply chain were explored via stochastic modeling with periodic review and order-up-to policy within distinct market structures [7]. Moreover, Cohen and Lee [8] developed a model, which integrated production, material management and distribution modules. They addressed decision-making via optimization on lot size, order-up-to quantity, and review period. A probabilistic model was developed to assess service and slack levels of a supply chain within lumpy demands [9]. By presenting a simulation study, real options strategies in the inventory management of strategic commodity-type parts in a supply chain are discussed. [10] Information distortion in a supply chain and its causes was investigated by [11].

These researches mainly focus on how to attack the issues regarding demand/supply uncertainty, and information incompleteness. Notably, each member of a supply chain possesses a certain degree of freedom with which to make decisions. One of the major concerns is thus how to cooperate among the autonomous members.

Within a supply chain, inventory ordering policies can be primarily classified into two categories, continuous review policy and periodic review policy. Basically, the former incorporates either a reorder point and order-up-to level policy, or a reorder point and order-quantity policy. The latter assumes that a periodic review time exists, and thus, considers order-up-to level policy and order-quantity policy, respectively.

Herein, we will focus on a serial supply chain using periodic review policy with order-up-to level. Although progressive work has been done by researchers to resolve

the issue of material management decision-making in a decentralized manner, how to develop an appropriate order-up-to level policy in a distributed manner for a serial supply chain are worthy to be further explored, such that the impact of demand/supply uncertainty and asymmetric information can be minimized.

3 The Serial Supply Chain Model

During each review period, a specific station examines demand requirements that stem from its downstream station. This review period results in information (in terms of local inventory level and refill rate) propagation delays within the supply chain. Besides, material transferring between stations requires a lead time that causes physical material delays. These delays disturb the demand-and-supply relation between two conjunctive stations, resulting in shortages, excess stock and backlogs. All the lead times are known and fixed to 1, excluding the first station (i.e., the retailer facing customers), which is zero. OUTL of a station i is denoted as S^i . Each station is characterized by its OUTL, transportation lead time, and inventory review period. It is also assumed that the supply chain produces a single product and has no capacity limitations. At the end of a review period, each station will examine its inventory and place orders from its upstream station as required. As well, backlog is allowed. In the highly distributed supply chain, each station is autonomous. It is assumed that refill rate and cost information is only shared between two conjunctive stations. In such an environment, a station will sacrifice local benefit-- refill rate and cost, to improve the refill rate for customers.

The supply chain in our following illustration contains several stations. The first is nearest to the customers and the last station is a raw material receiving station. The overall supply chain OUTL is represented by a vector \bar{S} .

Fuzzy membership function is applied to represent material supply and customer demand within an uncertain environment. Supplier has various 'reliability' modes—reliable and unreliable. The reliability of supply is represented by a membership function. As well, customer demand, per period, is denoted as a membership function.

Notations are defined as follows.

FR^i : The refill rate of station i . To represent a proportion of quantity to supply a downstream station ($i-1$). FR^1 is the customer service level of the supply chain.

For each station, let

d_j : Customer (or the downstream station) demand in period j .

I_K : Inventory in the end of K th review period, where

$$I_K = \begin{cases} P_{K-1} + I_{K-1} - D_K, & I_{K-1} - D_K \geq 0 \\ 0, & \text{elsewhere} \end{cases}, K = 1, 2, \dots, M, D_K = \sum_{j=1+(K-1)R}^{KR} d_j, P_K$$

is the replenishment quantity in K th review period. I_0 and P_0 equal to zero. R is the number of period within each review period.

S_K^i : Shortages at the end of review period K of station i , where

$$S_K = \begin{cases} -I_K, & \text{if } I_K < 0. \\ 0, & \text{if } I_K \geq 0 \end{cases}$$

$D = \sum_{j=1}^N d_j$: Total demand within N periods, where $N = MR$.

Hence, FR^i is obtained through $1 - \frac{\sum_{K=1}^M S_K^i}{D}$, $0 \leq FR^i \leq 1$, $i = 1, 2, \dots, 5$.

And the required replenishment quantity of a station from its upstream station is equal to OUTL of the station minus the sum of downstream station’s demand per review period (D_K) and inventory level (I_K).

Cost of a station is comprised of holding cost, shortage cost, and overhead cost owing to a change in OUTL, which requires that the warehouse be re-sizing as well as the material handling facility. These costs in the K th review period can be easily computed, and denoted as $C_h^i(K)$, $C_s^i(K)$ and C_a^i , respectively. The total holding, shortage, and OUTL overhead costs are defined as $\sum_{K=1}^M C_h^i(K)$, $\sum_{K=1}^M C_s^i(K)$ and $\sum_{K=1}^M C_a^i(K)$, respectively. The total cost of the i th station is therefore

$$TC^i = \sum_{K=1}^M C_h^i(K) + \sum_{K=1}^M C_s^i(K) + \sum_{K=1}^M C_a^i(K).$$

The average cost per product is defined as

$$\overline{TC} = \sum_{i=1}^N \overline{TC}^i \text{ where } \overline{TC}^i = \frac{TC^i}{D}.$$

Our research framework is as follows. Reliable material supply (100%) is simulated first. The resulting FR^1 is a benchmark service level (BSL). Then, refill rates and cost changes are examined when a material supplier’s reliability decreases. To increase the refill rate up to its BSL, a coordinated OUTL policy is introduced. To construct a simulation system for the serial supply chain, a variable-driven simulation tool, PowerSim [13], is employed. Batch mean method [14] is applied for output analysis and simulation stability is validated through a warm-up checking.

4 Coordinated Order-Up-To Level Policy

This study developed a coordinated order-up-to level inventory policy and its computational algorithm. In the serial chain, each station is autonomous and thus, decides its respective OUTL. Each station has its complete cost information and can also obtain partial information from its neighboring stations. For instance, the upstream station’s OUTL and current refill rate of its downstream station can be acquired. When a station adjusts its OUTL, the cost of the station will be evaluated. Finally, each station aims to minimize its local cost.

The rationale for the design of the OUTL algorithm can be described as follows. A poor refill rate within downstream station indicates the necessity to increase the OUTL of its corresponding upstream station. Similarly, a high OUTL within an upstream station implies a high OUTL of the downstream station.

SC is employed to evaluate the ‘current’ system performance, via formula 1, in transient stages of the proposed OUTL algorithm described below. The definition of SC indicates that a higher value is preferred. Considering multiple measures in supply chain has been widely investigated and reported [3]. The synthesized performance measure using refill rate and production costs is gauged to evaluate the OUTL policies. FR^1 and CI of formula (1) are normalized numbers, such that unit and scale are eliminated and different experimental settings can be fairly compared. $\Phi(\bullet)$ maps its content to a linear function between $[0, \overline{TC}_{goal}]$.

$$SC = x (FR^1) + (1-x) (CI) , \text{ where } x \in [0,1], \text{ and} \tag{1}$$

$$CI = 1 - \frac{\Phi(\overline{TC}_{current} - \overline{TC}_{goal})}{\overline{TC}_{goal}} , \quad \Phi(\overline{TC}_{current} - \overline{TC}_{goal}) = \begin{cases} 0 & , \text{ if } \overline{TC}_{current} - \overline{TC}_{goal} \leq 0 \\ \overline{TC}_{goal} & , \text{ if } \overline{TC}_{current} \geq 2\overline{TC}_{goal} \\ \overline{TC}_{current} - \overline{TC}_{goal} & , \text{ otherwise} \end{cases}$$

When material supply is reliable (100% supply), \overline{TC}_{goal} is the corresponding average cost per product. $\overline{TC}_{current}$ is the average cost per product under the current supply reliability.

SR, which resembles SC, is computed for reliable supply.

The procedure contains three stages: initialization stage, supply reliability changing stage, and each-station-executing-OUTL stage. In the initialization stage, BSL is measured for reliable supply. An initial OUTL \overline{S}_0 and SR are obtained. Stage two computes SC for unreliable supply.

Stage three triggers the processes to decide an appropriate amount for increasing OUTL, such that the serial chain can recover its performance from unreliable material supply. The procedure follows a standard hill-climbing and neighborhood search algorithm [15]. In Stage three, every station in the serial chain asynchronously executes the coordinated OUTL policy that follows seven steps to decide individual OUTLs.

Step I. If the inventory level of an upstream station has not increased, then halt; otherwise, go to Step II.

Step II. (denoted as D algorithm) Each station i negotiates with its neighboring upstream and downstream stations asynchronously. OUTL is adjusted with formula (2), where OUTL adjustment continues until $FR_p^i > FR_A^i$; that is each station constantly increases its OUTL by one unit before the refill rate under current (unreliable) supply environment reaches its original refill rate under reliable supply. The OUTL setting (x) with maximal K_p is selected.

$$\underset{x=S_0^i \text{ to } +\infty}{MAX} \{K_p(x) | K_p(x) = \omega_F \times (FR_p^i(x) - FR_0^i) + \omega_C \times (C_0^i - C_p^i(x)) \text{ given } B\} \quad (2)$$

where ω_F : weight of refill rate. ω_C : weight of cost, and $\omega_F + \omega_C = 1$. C_0^i : cost of station i when supply reliability reduces and a new OUTL is not applied. $C_p^i(x)$: cost of station i when supply reliability reduces and a new OUTL x is applied. FR_0^i : refill rate of station i when supply reliability reduces and a new OUTL is not applied. $FR_p^i(x)$: refill rate of station i when supply reliability is reduced and a new OUTL x is applied. FR_A^i : refill rate of station i when supply reliability is high. B is a conditional event in that when $FR_p^i \leq FR_A^i$, increase OUTL by one unit; otherwise, stop the heuristic at the station.

The purpose of step II is to coordinate the optimal OUTL of a station, such that refill rate to its succeeding station is thus maximized and its own cost is increased to the lowest level.

Step III. Compute SC and ϵ_0 , a proportion of SC and SR before the 1th iteration.

Step IV. Adjust ω_F to locate the local optimal solution between two adjacent stations. An increase in ω_F thus decreases ω_C of a station, indicating that the refill rate of its downstream station can be sacrificed. Similarly, the increase in ω_C , thus decreases ω_F , indicating that its downstream station can improve the refill rate.

Step V. Re-run D algorithm. Computes ϵ_m in the m th iteration to decide the search direction.

Step VI. Do a neighborhood random search of \bar{S} through adjusting ω_F by \bullet units (predefined to specify incremental precision of ω_F). If $\epsilon_m < \epsilon_0$, switches the search direction of \bar{S} in the opposite (when an improvement trial of SC through changing \bar{S} is failed); otherwise, search on the same direction.

Step VII. Execute D algorithm and compute ϵ_n until $\epsilon_n < \epsilon_{n-1}$ (an improvement trial of SC is failed). When

$\epsilon_n < \epsilon_{n-1}$, choose the OUTL \bar{S} of the $n-1$ th iteration, and stop the procedure.

For the clarifying the proposed procedure, a simulation model via PowerSim simulation package mimicking a five-station serial chain is designed to illustrate the proposed policy.

Table 1. Simulation setting

Factor	Value/Definition
Supply reliability (μ_r)	Fuzzy membership function: Reliable{100% / 1}; Unreliable{80% / 1, 90% / 0.5, 100% / 0.25}
Initial OUTL (\bar{S})	(36,36,36,36,36)
Lead time (L^i)	1
Review period (R)	4
Unit holding cost (C_h^i)	$C_h^1 = 1.5, C_h^2 = 1, C_h^3 = 0.5, C_h^4 = 0.3, C_h^5 = 0.1$
Unit shortage cost (C_s^i)	$C_s^1 = 6.5, C_s^2 = 4.3, C_s^3 = 2.15, C_s^4 = 1.29, C_s^5 = 0.43$
OUTL overhead cost (C_a^5)	$C_a^1 = 1.1, C_a^2 = 0.9, C_a^3 = 0.7, C_a^4 = 0.6, C_a^5 = 0.4$
Demand (μ_{DR})	{7/0.25, 8/0.5, 9/0.75, 10/1, 11/0.75, 12/0.5, 13/0.25}
Weight (x)	0.5

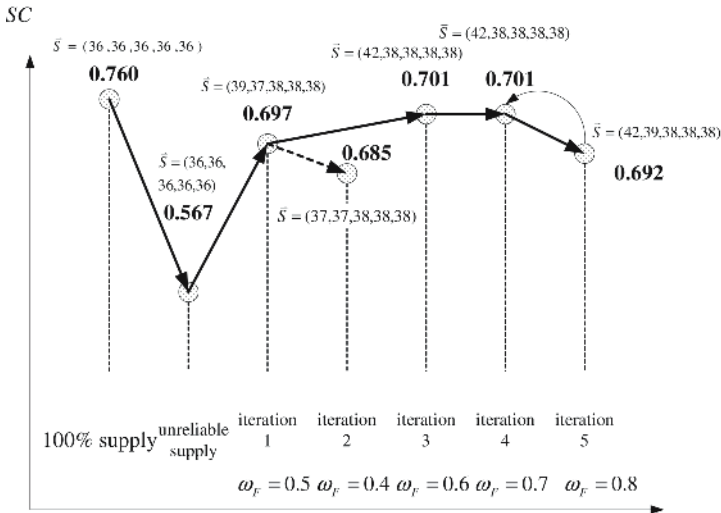


Fig. 1. Adjustment process

Setting of the number of stations to five is complex enough for observation and manageable for explanation. Table 1 displays the required parameter setting. An initial OUTL is established via a simulated defuzzifying process. When supply reliability reduces to a worse level, the supply chain is affected in which the *SC* value decreases from 0.760 to 0.567.

Figure 1 displays the process of the adjustment. Table 2 compares the supply chain performances of the proposed coordinated OUTL policy with those reliable and unreliable supply. The *SC* value increases from 0.567 to 0.701.

Table 2. A comparison of the supply chain performances.

Material supply reliability mode	Reliable	Unreliable (unadjusted \bar{S})	Unreliable (adjusted \bar{S})
<i>SC</i>	0.760 (=SR)	0.567	0.701
FR^1	0.52	0.35	0.82
\overline{TC}	7.513	9.131	10.659

5 Mediation Procedure

In the above procedure, it was ascertained that each station decides its own OUTL individually. If a mediator existed within the supply chain to coordinate OUTL, an improved solution could be attained. Therefore, to further improve supply chain performances, a mediation procedure was constructed.

In this mediation heuristic, when OUTL is adjusted, FR^1 and \overline{TC} are both considered. FR^1 and OUTL correlate positively. That is, in most cases, if OUTL is increased then FR^1 is improved. Alternately, \overline{TC} and OUTL represent a convex cost trade-off relation, which is due to both an excess and a shortage of materials. The mediation procedure is described as follows:

1. When (i) \overline{TC} exceeds C_{goal} twice (a pre-defined threshold value) over in reliable supply (7.513 in the case) or (ii) FR^1 exceeds *BSL*, the mediator requests that the station with the highest OUTL reduce its OUTL by one unit. When two or more stations have equal, highest OUTL, the one nearest the customer is reduced first.
2. The new \bar{S} is applied to system adjustment procedure (Section 4.1) and SC_{n+1} is computed until $SC_{n+1} < SC_n$. Select OUTL of the *n*-1th iteration.

Applying the mediation heuristic within unreliable supply, its performances are revealed. Prior to executing the mediation procedure of the unreliable supply case, the local optima is established at $\bar{S} = (42,38,38,38,38)$ and $SC_0 = 0.701$. The resulting *SC* is 0.716.

6 Performance Evaluation of the Coordinated OUTL Policy

The coordinated OUTL policy is compared to two OUTL heuristics: the single station and the multiple station policies (a detail description of the two heuristics refers to [4]). The single-station policy adjusts OUTL for only one arbitrary station. The multiple station policy adjusts the OUTL of the corresponding stations by using an evaluation index, $MF = \frac{FR^1 \times \Delta FR^1}{TC}$.

Note that OUTL overhead cost and the synthesis performance index (formula 1) are applied in all the three policies such that they are fairly compared. Simulation experiments setting is listed in table 1. Figure 1 indicates that the coordinated OUTL policy with the mediation procedure outperforms the other policies.

The robustness of the coordinated OUTL policy was examined based on a variety of x varying from 0.1 to 0.9 to represent a variety of viewpoints of performance focus (on refill rate or cost). Within all cases, the coordinated policy outperforms the others.

Three types of demand patterns represented as fuzzy member functions and are employed to test the three OUTL policies. Table 3 presents that the coordinated policy outperforms the others under all demand situations.

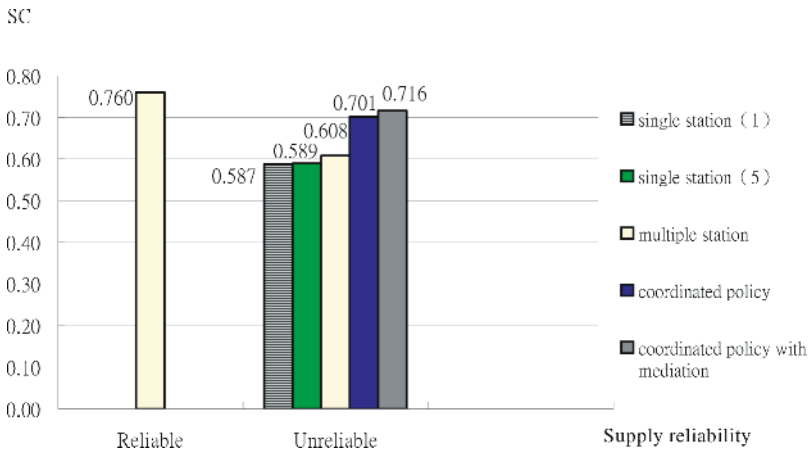


Fig. 2. Comparison of coordinated, single-station, and multiple-station policies

Table 3. The performances for different demand patterns under unreliable supply. Reliable (mean =10, Fuzzy member function= {10/1}); Unreliable(mean =10, Fuzzy member function = {7/0.25, 8/0.5, 9/0.75, 10/1, 11/0.75, 12/0.5,13/0.25})

Demand pattern	Single station policy applied to the 1 st station	Single station policy applied to the 5 th station	Multiple station policy	Coordinated policy with mediation procedure
Reliable	0.620	0.610	0.625	0.716
Unreliable	0.597	0.584	0.607	0.706

The experimental outcomes in the preliminary performance study of the proposed OUTL policy indicate that (i) the coordinated OUTL policy outperforms the single-station and multiple-station policies for a variety of viewpoints of performance focus either on refill rate or cost, and different demand patterns; and (ii) the mediation procedure is able to further improve supply chain performances on the basis of the coordinated OUTL policy.

7 Concluding Remarks

This study has proposed a method for coordinating the inventory decisions of isolated stations in a supply chain to cooperatively compensate loss of total cost and refill rate due to unreliable raw material supply. The proposed method consists of a coordinated OUTL policy and a mediation procedure. After compared to single-station and multiple-station policies using simulation and through a comprehensively experimental study, it indicates that the proposed method outperforms the others from a variety of viewpoints of performance and under different demand patterns when material supply is unreliable.

References

1. Parker, R. P., and Kapuscinski, R. Managing a non-cooperative supply chain with limited capacity, Yale School of Management. <http://welch.som.yale.edu/researchpapers/> (2001)
2. Thirumalai, R. *Multi criteria - multi decision maker inventory models for serial supply chains*, PhD dissertation, Industrial Engineering, Penn State University (2001)
3. Beamon, B. M. Supply chain design and analysis: models and methods, *International Journal of Production Economics*, 55, (1998) 281-294
4. Petrovic, D., Roy, R., and Petrovic, R. Modeling and simulation of a supply chain in an uncertain environment, *European Journal of Operational Research*, 109, (1998) 299-309
5. Petrovic, D., Roy, R., and Petrovic, R. Supply chain modeling using fuzzy set, *International Journal of Production Economics*, 59, (1999) 443-453
6. Lee, H. L., Billington, C., Carter, B. Hewlett-Packard gains control of Inventory and Service Through Design for Localization, *Interfaces*, 23(4), (1993) 1-11
7. Lee, H. L. and Billington, C. Material management in decentralized supply Chains, *Operations Research*, 41(5), (1993) 835-847
8. Cohen, M. A. and Lee, H. L. Strategic analysis of integrated production-distribution systems: Models and methods, *Operations Research*, 36(2), (1998) 216-228
9. Verganti, R. Order overplanning with uncertain lumpy demand: a simplified theory, *International journal of production research*, 35(12), (1997) 3229-3248
10. Marqueza, A. C. and Blancharb, C. The procurement of strategic parts. Analysis of a portfolio of contracts with suppliers using a system dynamics simulation model, *Int. J. Production Economics*, 88 (2004) 29-49.
11. Lee, H. L., Padmanabhan, V., Whang S. Information distortion in a supply chain: the bullwhip effect, *Management Science*, 43(4), (1997) 546-558
12. Zimmermann, H. J. *Fuzzy set theory-and its applications*, Kluwer Academic Publishers (1991)
13. PowerSim Corporation PowerSim, *Reference Manual* (2005)
14. Law, A. M. and Kelton, W. D. *Simulation modeling and analysis*, McGraw-Hill, Inc., New York (1991)
15. Russell S. and Norvig P. *Artificial intelligence: a modern approach*, Prentice Hall International (2003)

Optimization of Performance of Genetic Algorithm for 0-1 Knapsack Problems Using Taguchi Method

A.S. Anagun and T. Sarac

Eskisehir Osmangazi University, Industrial Engineering Department,
Bademlik 26030, Eskisehir, Turkey
{sanagun, tsarac}@ogu.edu.tr

Abstract. In this paper, a genetic algorithm (GA) is developed for solving 0-1 knapsack problems (KPs) and performance of the GA is optimized using Taguchi method (TM). In addition to population size, crossover rate, and mutation rate, three types of crossover operators and three types of reproduction operators are taken into account for solving different 0-1 KPs, each has differently configured in terms of size of the problem and the correlation among weights and profits of items. Three sizes and three types of instances are generated for 0-1 KPs and optimal values of the genetic operators for different types of instances are investigated by using TM. We discussed not only how to determine the significantly effective parameters for GA developed for 0-1 KPs using TM, but also trace how the optimum values of the parameters vary regarding to the structure of the problem.

1 Introduction

The KP is a well-known combinatorial optimization problem. The classical KP seeks to select, from a finite set of items, the subset, which maximizes a linear function of the items chosen, subject to a single inequality constraint.

KPs have been mostly studied attracting both theorists and practitioners. 0-1 KP is the most important KP and one of the most intensively studied discrete programming problems. The reason for such interest basically derives from three facts: (a) it can be viewed as the simplest integer linear programming problem; (b) it appears as a sub problem in many more complex problems; (c) it may represent a great many practical situations [1].

Many optimization problems are combinatorial in nature as 0-1 KP and quite hard to solve by conventional optimization techniques. Recently, GAs have received considerable attention regarding their potential as an optimization technique for combinatorial optimization problems [2-5].

In this paper, we present an approach for determining effective operators of GAs (called design factors) for solving KP and selecting the optimal values of the design factors. The KP is defined as follows:

$$\max_x \left\{ \sum_{j=1}^n p_j x_j \mid \sum_{j=1}^n w_j x_j \leq c, x \in [0,1]^n \right\} \quad (1)$$

where n items to pack in some knapsack of capacity c . Each item j has a profit p_j and weight w_j , and the objective is maximizing the profit sum of the included items without having the weight sum to exceed c .

Since the efficiency of GAs depends greatly on the design factors, such as; the population size, the probability of crossover and the probability of mutation [6], selecting the proper values of the factors set is crucial. Various reproduction and crossover types can be used in regard to optimizing the performance of the GA. In addition, operator types may affect the efficiency of GA. In this study, to investigate the affect of the factors briefly discussed, we generated nine instances composed of three different sizes (50, 200 and 1000) and three different types correlation (*uncorrelated*, *weakly correlated* and *strongly correlated*). To optimize the efficiency of GAs in solving a 0-1 KP, we examined closely which genetic operators must be selected for each instance by using TM.

The paper is organized in the following manner. After giving brief information about 0-1 KP, the GA for 0-1 KPs and the parameters that affect the performance of the GA are introduced. The fundamental information about TM is described next. Then the steps of experimental design are explained and applied to the problems taken into account. The results are organized regarding to the problems. Finally, the paper concludes with a summary of this study.

2 Genetic Algorithm for 0-1 Knapsack Problems

GAs are powerful and broadly applicable in stochastic search and optimization techniques based on principles from evolution theory [7]. GAs, which are different from normal optimization and search procedures: (a) work with a coding of the parameter set, not the parameters themselves. (b) search from population of points, not a single point. (c) use payoff (objective function) information, not derivatives or other auxiliary knowledge. (d) use probabilistic transition rules, not deterministic rules [8].

KPs that are combinatorial optimization problems belong to NP-hard type problems. An efficient search heuristic will be useful for tackling such a problem. In this study, we developed a GA (KP-GA) working with different crossover operators (1:*single-point*, 2:*double-point*, 3:*uniform*) and reproduction types (1:*roulette wheel*, 2:*stochastic sampling*, 3:*tournament*) to solve the 0-1 KP. The KP-GA is coded with VBA. The proposed GA for solving 0-1 KPs is discussed below.

2.1 Coding

We shall use an n bit binary string which is a natural representation of solutions to the 0-1 KPs where one means the inclusion and zero the exclusion of one of the n items from the knapsack. For example, a solution for the 7-item problem can be represented as the following bit string: [0101001]. It means that items 2, 4, 7 are selected to be filled in the knapsack. This representation may yield an infeasible solution.

2.2 Fitness

The fitness value of the string is equal to the total of the selected items profit. To eliminate the infeasible string we use penalty method. If any string is not feasible, its fitness value is calculated by using the procedure penalty is given below:


```

procedure penalty:
  if  $\sum_{i=1}^n w_i > c$  then;
    begin
      p :=  $1 - (\sum_{i=1}^n w_i / 5 * c)$ ;
      if p <= 0 then p := 0.00001;
      fitness value of string i = fitness value of string i * p;
    end;

```

2.3 Genetic Operators

A classical GA is composed of three operators: reproduction, crossover and mutation. Operators taken into consideration for solving 0-1 KP problems with GA are discussed as follows.

The reproduction operator allows individual strings to be copied for possible inclusion in the next generation. The chance that a string will be copied is based on the string's fitness value, calculated from a fitness function. We use three types of reproduction operators; roulette wheel, remainder stochastic sampling, and tournament.

Crossover enables the algorithm to extract the best genes from different individuals and recombine them into potentially superior children. KP-GA has three different crossover operators; single point, double point and uniform.

Reproduction and crossover alone can obviously generate a staggering amount of differing strings. However, depending on the initial population chosen, there may not be enough variety of strings to ensure the GA searches the entire problem space, or the GA may find itself converging on strings that are not quite close to the optimum it seeks due to a bad initial population. Some of these problems may be prevented by introducing a mutation operator into the GA. In KP-GA, a random number is generated for all genes. If the random number is smaller than mutation rate, the value of the gene is changed. If it is 0, its new value will be 1 and if it is 1, it will be 0, respectively. The value of gene is protected, if random number is bigger than the mutation rate.

When creating a new generation, there is always a risk of losing the most fit individuals. Using elitism, the most fit individuals are copied to the next generation. The other ones undergo the crossover and mutation. The elitism selection improves the efficiency of a GA considerably, as it prevents losing the best results.

The termination condition is a check whether the algorithm has run for a fixed number of generations. The number of generations of 1000 is selected as termination condition for all of the experiments conducted.

3 Taguchi Method

In principle, Taguchi design of experiments is used to get information such as main effects and interaction effects of design parameters from a minimum number of experiments. The behavior of a product or process is characterized in terms of design (*controllable*) and noise (*uncontrollable*) factors. Thus, TM may be applied to determine the best combination of design factors and to reduce the variation caused by the noise factors [9].

The TM uses a special design of orthogonal arrays (OAs) to study the entire parameter space with only a small number of experiments. Experimental design using OAs, recommended by Taguchi, not only minimizes the number of treatments for each trial, but also keeps the pair-wise balancing property [10].

An OA is basically a matrix of rows and columns in which columns are assigned to factors or their interactions and rows represent the levels of various factors for a particular experimental trial [11]. The treatment combinations are chosen to provide sufficient information to determine the factors' effects using the analysis of means. The OA imposes an order on the way the experiment is carried out. Orthogonal refers to the balance of the various combinations of factors so that one factor is given more or less weight in the experiment than the other factors. Orthogonal also refers to the fact that the effect of each factor can be mathematically assessed independently of the effects of the other factors [12].

The first step of designing an experiment with known numbers of factors in Taguchi's method is to select a most suitable OA, which design to cover all the possible experiment conditions and factor combination. The selection of which OA to use predominantly depends on these items in order of priority [13]: (1) the number of factors and interactions of interest, (2) the number of levels for the factors of interest, and (3) the desired experimental resolution or cost limitations.

In order to select an appropriate OA for experiments, the total degrees of freedom need to be computed. The degrees of freedom are defined as the number of comparisons between design factors that need to be made to determine which level is better and specifically how much better it is [14]. While the degrees of freedom associated with a design factor is one less than the number of levels, the degrees of freedom associated with interaction between two design factors is given by the product of the degrees of freedom for the two design factors. Basically, the degrees of freedom for the OA should be greater than or at least equal to those for the design factors.

Once the levels of design factors are settled, the analysis of means (ANOM) is conducted to find affection of each factor on the performance criterion by calculating the mean of entire data of the design factors. Hence, the optimum level of each design factor can be found by concentrating on its corresponding response graph. The analysis of variance (ANOVA) is then performed to determine the significant factors for the selected criterion. Finally, a prediction model consisting of the significant factors is built and confidence intervals for estimated mean and each of the significant factors are constructed.

4 The Experimental Design

In this paper, TM is applied to search the optimum values of the parameters of GAs designed for 0-1 KPs which are generated in terms of the number of items and correlation among weights and profits of items, respectively. With this consideration, it is aimed that whether the optimum values of the parameters affecting the performance of GAs may be changed while the numbers of items and correlation among weights and profits of items increasing and/or decreasing.

Five design factors are identified as potentially important for performance of GAs: (1) population size, (2) crossover rate, (3) mutation rate, (4) crossover type, and (5)

reproduction type. For each of the design factors, based on the related research, three possible levels were considered as: population size of (10, 30, 50), crossover rate of (0.60, 0.75, 0.90), mutation rate of (0.001, 0.005, 0.01), crossover type (1, 2, 3) and reproduction type of (1, 2, 3), respectively.

L_{27} (3^{13}) OA was selected since it is the most suitable plan for the conditions being investigated, which allows for examining 13 three-level design factors and/or interactions among them with 27 trials. Regarding to the light of the preliminary tests, the selected factors and the interactions between design factors were assigned to the columns of the OA before conducting tests. For further information on OAs and assigning factors to an OA, readers may refer to [9].

In order to observe the effects of noise sources, each experiment was repeated three times ($3 \times 27 = 81$) under the same conditions. The order to the experiments was made random in order to avoid noise sources that had not been taken into account initially and that could affect the results in a negative way. To investigate how KP-GAs behave for different 0-1 KPs, nine types of data instances were randomly generated regarding to the number of items and the correlation among weights and profits of items. The L_{27} OA was then applied to each of the KPs and the optimum levels of the design factors that affect the performance of the KP-GA were determined separately.

5 Results

Based on the combinations of design factors assigned to the selected OA, twenty-seven different KP-GAs are formed for each of the 0-1 KPs. Three types of instances representing the correlation among weights and profits of items are generated by using the following definitions and Eq. (2) that are proposed by [15]:

Uncorrelated instance: the weights w_j and the profits p_j are uniformly random distributed in $[1, R]$, $R=1000$.

Weakly correlated instance: the weight w_j are distributed in $[1, R]$ and the profits p_j in $[w_j - R/10, w_j + R/10]$ such that $p_j \cdot 1$.

Strongly correlated instance: the weights w_j are distributed in $[1, R]$ and the profits are set to $p_j = w_j + R/10$.

$$\text{Capacity is chosen as } c = \frac{1}{10} \sum_{j=1}^n w_j \quad (2)$$

The data obtained from the trials; for instance, when the number of items is 50 and there is no correlation among instances, coded as [50_Unc], are analyzed as follows.

5.1 Analysis of Mean (ANOM)

As mentioned in [16], the Taguchi has created the S/N (signal to noise) ratio to quantify the present variation. The term “*signal*” represents the desirable value (mean) and the term “*noise*” represents the undesirable value (standard deviation). There are several S/N ratios depending on the types of characteristics; lower the better (LB), nominal the best (NB), and higher the better (HB). Since the nature of the objective function for KPs is maximization, higher the better (HB) criterion was selected as a performance statistics:

$$Z_{HB} = -10 \log \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{y_i^2} \right] \tag{3}$$

where n is the number of repetition of simulation under the same condition of design factors, y the characteristics, and subscript i indicates the simulation number of design factors in the OA table. As shown in Eq. (3), the greater the S/N ratio, the smaller is the variance of performance measure around the desired value.

The data for [50_Unc] and Eq. (3) are applied to obtain the S/N response graph which is depicted in Fig.1. From Fig.1, the optimum levels of the design factors of KP-GA designed for [50_Unc], A3B2C3D3E1 can be found and its corresponding values are shown as; A3: population size of 50, B2: crossover rate of 0.75, C3: mutation rate of 0.01, D3: crossover type of 3, and E1: reproduction type of 1.

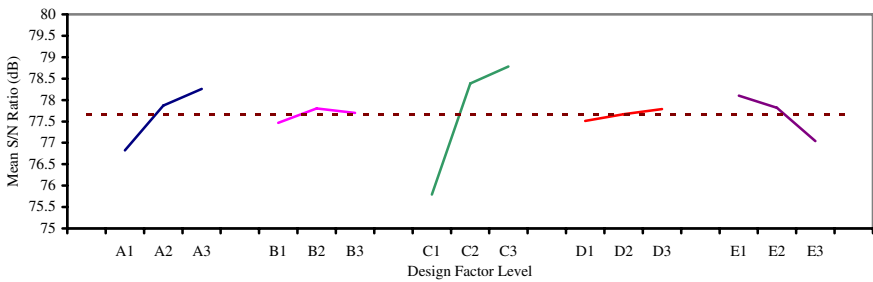


Fig. 1. S/N response graph for [50_Unc]

5.2 Analysis of Variance (ANOVA)

ANOVA, one of the most commonly and widely used analytical tools, is a technique that subdivides the total variation into meaningful components or sources of variation [17]. Using the ANOVA, one may be able to spot the key components (or factors) that cause excessive variation in products or processes. The ANOVA can be done with the raw data or with the S/N data.

The ANOVA based on the raw data signified the factors which affect the average response rather than reducing the variation. On the other hand, the ANOVA based on the S/N data takes into account both these aspects and so it was used here. Since all columns in OA are assigned with the factors and interactions, called saturated design, the variations due to error are estimated by pooling the estimates of the factors and interactions having least variance. This also helps in determining Fisher test (F-test) for finding out the confidence level of the results.

The ANOVA performed for the S/N data of [50_Unc] using ANOVA_TM software is given in Table 1. Table 1 indicates that factor C has the largest effect on the performance criterion. Factor A and factor E have the next largest effects, respectively. The remaining design factors have no effects due to their F values. Table 1 also indicates that the interaction of AxC has moderate effect and the interaction of CxE has least effect on the performance criterion. It is observed that the response graphs

for the interactions' effects provided the identical results as the ones obtained regarding to the response graphs for the design factors. However, the interactions, in addition to the main design factors, should be considered to calculate the estimated mean S/N for the problem of [50_Unc].

Table 1. ANOVA for S/N data of [50_Unc] – After pooling

Source	Pool	Df	S	V	F	S'	rho%
A	[N]	2	9.92062	4.96031	28.00694	9.56640	11.68
B	[N]	2	0.53491	0.26745	1.51008	0.18169	0.22
CxD	[Y]	4	0.56962	0.14240	-	-	-
CxE	[N]	4	2.26365	0.56591	3.19525	1.55521	1.90
C	[N]	2	47.46478	23.73239	133.99802	47.11056	57.50
AxC	[N]	4	14.52187	3.63047	20.49839	13.81343	16.86
BxC	[Y]	4	0.85650	0.21412	-	-	-
D	[Y]	2	0.34502	0.17251	-	-	-
E	[N]	2	5.45266	2.72633	15.39343	5.09844	6.22
(e)		10	1.77114	0.17711		4.60490	5.62
Total	[-]	26	81.92963	3.15114			

The column under the rho% gives an idea about the degrees of contribution of the factors to the performance criterion. To verify that the factors and levels selected for the experiment are reasonably correct, after pooling, the rho% accounted by the different factors cumulatively should be greater than 60% [18]. Since the total of rho% regarding to the significant design factors and interactions is approximately 94.4%, it may be said that the conducted experiment to find the optimal values of the design parameters for KP-GA applied to [50_Unc] problem are appropriate and furthermore, confirmation tests may be run.

5.3 Confirmation Tests

In order to predict and/or verify the improvement of the performance characteristic, confirmation tests should be conducted using the optimal levels of the design factors based on the results obtained from ANOM and ANOVA. The estimated S/N ratio η_{opt} using the optimal levels of the main design factors and interactions can be calculated as [19]:

$$\eta_{opt} = \hat{\eta} + \sum_{j=1}^o (\bar{\eta}_j - \hat{\eta}) \tag{4}$$

where $\hat{\eta}$ is the total mean S/N ratio, $\bar{\eta}_j$ the mean S/N ratio of the main design factor or interaction at the optimal level, and o the number of the design factors that affect the performance characteristic. In regard to optimal levels of the design factors that significantly affect the performance of KP-GA used for [50_Unc], the estimated S/N ratio of 78.94190 (dB) is computed using Eq. (4).

The confidence interval of the estimated mean S/N ratio can be calculated by considering the following equation to verify whether the optimal solution of the objective function for the problem of [50_Unc] or targeted value is reached [20]:

$$CI = \sqrt{\frac{F(\alpha,1,v_e) V_e}{n_{eff}}}; n_{eff} = \frac{N}{1 + \sum(\cdot)} \tag{5}$$

where $F(\alpha,1,v_e)$ is the F-ratio required for $\alpha = \text{risk}$, v_e the degrees of freedom of error, V_e the pooled error variance, n_{eff} the effective sample size, N is the total number of trials, and $\sum(\cdot)$ the total degrees of freedom associated with items used in η_{opt} estimate. A confidence interval of 95% for the estimated S/N, $F(0.05,1,10) = 4.96$, $V_e = 0.17711$, and the effective sample size is $n_{eff} = 1.8$. Thus, the 95% confidence interval of the estimated optimum (dB) is computed as [78.24330; 79.64049].

The optimum levels of the design factors obtained for the problem of [50_Unc] are applied to the other 0-1 KPs generated with respect to the number of items and the correlation among weights and profits of items to examine whether the levels determined for the problem of [50_Unc] by means of TM are appropriate for the remaining generated 0-1 KPs. It is observed that the optimum levels of the design factors are basically problem dependent meaning that the appropriate levels of the design factors should be determined for each of the 0-1 KPs separately. Hence, each of the problems generated are analyzed based on the steps given in Section 5. In addition to results of problem [50_Unc], the results obtained from analyses for the remaining 0-1 KPs being solved by KP-GA are shown in Table 2.

As shown in Table 2, the estimated S/N means for the generated 0-1 KPs are somehow different than the optimal solution obtained using LINGO, software for optimization modeling, although the constructed confidence levels included the optimal solutions for each of the problems concerned.

For a validity check, all of the experiments are repeated with the best combinations by considering different number of generations to verify whether the optimal solutions for the problems may be obtained. These experimentations concluded that the best combinations determined by the TM are able to produce the solutions as close as the ones obtained using LINGO, which makes the TM a powerful and sensitive approach for solving 0-1 KPs even if the size of the problems vary in terms of the numbers of items and the correlation among weights and profits of items.

Table 2. Results for analyses of generated 0-1 KPs

Problem Type	Treatment Combination	Estimated S/N Mean	Confidence Interval (95%)	Converted S/N*	
Number of items	Correlation				
50	No	A3B2C3D3E1	78.94190	[78.24330; 79.64049]	78.9271
50	Weak	A3B3C3D1E1	68.97514	[68.70190; 69.24840]	68.9307
50	Strong	A3B3C3D3E1	72.02041	[71.73480; 72.30610]	71.9209
200	No	A2B1C2D1E3	90.53733	[90.38260; 90.69206]	90.6665
200	Weak	A3B3C3D1E2	81.39253	[81.15990; 81.62520]	81.6133
200	Strong	A3B3C3D3E2	83.96448	[83.65620; 84.27280]	83.9770
1000	No	A3B1C1D1E2	104.12979	[103.2520; 105.0076]	105.0733
1000	Weak	A3B1C1D1E2	95.11788	[94.93680; 95.29900]	96.0850
1000	Strong	A3B1C2D1E2	97.53570	[97.36880; 97.70260]	98.2418

(*) optimal solution of the 0-1 KPs obtained from LINGO.

6 Conclusion

The aim of this study is not only to determine the optimum parameters that affect the performance of a GA designed for a combinatorial optimization problem, but also to investigate whether or not the optimum parameters of a GA applied to the problems, specifically 0-1 KPs formed in terms of the numbers of the items and the correlation among weights and profits of items, vary. The results obtained from the experiments conducted are summarized below.

The TM may be used to designate appropriate combinations of the parameters of the GA such that the optimal solutions may be reached; exactly the same solution for small and medium size problems, as close as the solution for the large size problems regardless of the correlation among weights and profits of items, comparing with the optimal solution obtained using LINGO.

The optimum value of 50 is obtained for the population size regardless of the structures of the problems concerned. The mutation rate is decreased as the size of the problem increased.

The double point crossover operator is not significant for the 0-1 KPs generated with respect to items and correlation among weights and profits of items. On the other hand, the appropriate crossover operator is obtained as uniform for small size problems, while single point crossover operator for large size problems despite the correlation.

The stochastic sampling reproduction operator seems appropriate for medium and large size problems, while roulette wheel is suitable for small size problems. Since the combinations of the parameters change as the correlation among weights and profits of items shift, as well as the number of items, the correlation should be taken into consideration for solving such problems.

The optimum parameters of GA designed for solving 0-1 KPs are dependent on the structure of problem. Hence, the initial values of the parameters may be selected using the results of this study as a table look-up. In order to provide a general outline for the persons who are interested in solving such problems by means of GA, the GA should be run for the different problems regarding to the numbers of items than the ones used in this study.

References

1. Martello, S., Toth, P.: *Knapsack Problems Algorithms and Computer Implementations*. John Wiley&Sons, England (1990)
2. Sakawa, M., Kato, K.: Genetic Algorithms with Double Strings for 0-1 Programming Problems. *European Journal of Operational Research*. 144 (2003) 581-597
3. Bortfeldt, A., Gehring, H.: A Hybrid Genetic Algorithm for the Container Loading Problem. *European Journal of Operational Research*. 131 (2001) 143-161
4. Taguchi, T., Yokota, T.: Optimal Design Problem of System Reliability with Interval Coefficient Using Improved Genetic Algorithms. *Computers and Industrial Engineering*. 37 (1999) 145-149
5. Michelcic, S., Slivnik, T., Vilfan, B.: The Optimal Cut of Sheet Metal Belts into Pieces of Given Dimensions. *Engineering Structures*. 19 (1997) 1043-1049
6. Iyer, S.K., Saxena, B.: Improved Genetic Algorithm for the Permutation Flowshop Scheduling Problem. *Computers and Operations Research*. 31 (2004) 593-606

7. Gen, M., Cheng, R.: Genetic Algorithms and Engineering Design. John Wiley&Sons, New York (1997)
8. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Reading (1989)
9. Taguchi, G.: Systems of Experimental Design. Unipub Kraus International Publishers, New York (1987)
10. Georgilakis, P., Hatzargyriou, N., Paparigas, D., Elefsiniotis, S.: Effective Use of Magnetic Materials in Transformer Manufacturing. *Journal of Materials Processing Technology*. 108 (2001) 209-212
11. Antony, J., Roy, R.K.: Improving the Process Quality Using Statistical Design of Experiments: A Case Study. *Quality Assurance*. 6 (1999) 87-95
12. Fowlkes, W.Y., Creveling, C.M.: Engineering Methods for Robust Product Design. Addison-Wesley, Reading (1995)
13. Ross, P.J.: Taguchi Techniques for Quality Engineering. 2nd edn. McGraw-Hill, New York (1996)
14. Nian, C.Y., Yang, W.H., Tarng, Y.S.: Optimization of Turning Operations with Multiple Performance Characteristics. *Journal of Materials Processing Technology*. 95 (1999) 90-96
15. Martello, S., Pisinger, D., Toth, P.: New Trends in Exact Algorithms for the 0-1 Knapsack Problem. *European Journal of Operational Research*. 123 (2000) 325-332
16. Wu, D.R., Tsai, Y.J., Yen, Y.T.: Robust Design of Quartz Crystal Microbalance Using Finite Element and Taguchi Method. *Sensors and Actuators*. B92 (2003) 337-344
17. Roy, R.K.: A Primer on the Taguchi Method. VNR Publishers, New York (1999)
18. Ozalp, A., Anagun, A.S.: Analyzing Performance of Artificial Neural Networks by Taguchi Method: Forecasting Stock Market Prices. *Journal of Statistical Research*. 2 (2003) 29-45
19. Lin, T.R.: Experimental Design and Performance Analysis of TiN-coated Carbide Tool in Face Milling Stainless Steel. *Journal of Materials Processing Technology*. 127 (2002) 1-7
20. Syrcos, G.P.: Die Casting Process Optimization Using Taguchi Methods. *Journal of Materials Processing Technology*. 135 (2003) 68-74

Truck Dock Assignment Problem with Time Windows and Capacity Constraint in Transshipment Network Through Crossdocks

Andrew Lim^{1,2}, Hong Ma¹, and Zhaowei Miao^{1,2}

¹ Dept of Industrial Engineering and Logistics Management,
Hong Kong Univ of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

² School of Computer Science & Engineering,
South China University of Technology, Guang Dong, P.R. China

Abstract. In this paper, we consider the over-constrained truck dock assignment problem with time windows and capacity constraint in transshipment network through crossdocks where the number of trucks exceeds the number of docks available, the capacity of the crossdock is limited, and where the objective is to minimize the total shipping distances. The problem is first formulated as an Integer Programming (IP) model, and then we propose a Tabu Search (TS) and a Genetic algorithms (GA) that utilize the IP constraints. Computational results are provided, showing that the heuristics perform better than the CPLEX Solver in both small-scale and large-scale test sets. Therefore, we conclude that the heuristic search approaches are efficient for the truck dock assignment problem.

Keywords: Heuristic search, crossdocks, dock assignment.

Areas: Heuristics, industrial applications of AI.

1 Introduction

Dock assignment for trucks is one of key activities at crossdocks. Trucks are assigned to docks for the duration of a time period during which the cargo and trucks are processed. Dock availability and times of arrivals/departures (as given by an estimated time of arrival/departure or ETA/ETD for each truck) can change during the course of the planning horizon due to operational contingencies (for example, delays, traffic control). A familiar scene at crossdocks these days is when arriving trucks are waiting for process, sometimes for a long time, before finally proceeding to their docks, because the gate is occupied by another truck. So they need to schedule those docks well in order to increase the utilization and achieve better performance of the transshipment network. Firstly, good dock assignment can help crossdock increase the utilization by reducing dock delays. Secondly, good dock assignment can minimize distances (times) cargo are required to transfer from dock to dock. Because of the large number of freight and the dynamic nature of the problem, scheduling has become more difficult. This has made it more important for crossdock operators to use docks in the best possible way.

We consider the truck dock assignment in transshipment network through crossdocks. But in classical models, where transshipment is studied in the context of network flow [1]. One such model where transshipment becomes an important factor is in crossdocking which has become synonymous with rapid consolidation and processing. Tsui and Chang used a bilinear program of assigning trailers to doors, where the objective was to minimize weighted distances between incoming and outgoing trailers [8]. Recently, a study by Bartholdi and Gue examined minimizing labor costs in freight terminals by properly assigning incoming and outgoing trailers to doors [2]. Although previous cross-docking studies have considered intra-terminal factors such as types of congestion that impact costs, they do not address actual dock assignments to arriving vehicles when considering the time window of trucks and capacity of crossdocks.

Also our problem is similar in some ways to the problem of gate assignments in airports, for which some analytical work exists. For example, the basic gate assignment problem is a quadratic assignment problem and shown to be NP-hard [7]. Since the gate assignment problem is NP-hard, various heuristic approaches have been used by researchers and work has focused on the over-constrained airport gate assignment, where there is an excess of flights over gates [3, 6]. The objective there was to minimize the number of flights without any gate assigned (i.e. those left on the ramp) and the total walking distance.

In this paper, we consider the over-constrained truck dock assignment problem with time windows and capacity constraint in transshipment network through crossdocks where the number of trucks exceeds the number of docks available and the capacity of the crossdock is limited, and where the objectives are to minimize the total shipping distances. The problem is formulated as an IP problem. The air gate assignment problem is NP-hard, then our problem is also NP-hard because the air gate assignment problem is a special case of our problem. We use both a Tabu Search and a GA algorithms to solve the problem. Computational results are provided, showing that our heuristics work well in all the test sets.

This work is organized as follows: in the next section, we give an IP model. Tabu search and GA algorithms are developed in Section 3 and section 4. We provide computational results in Section 5. In Section 6, we summarize our findings.

2 Problem Description and Formulation

In this section, we provide an IP model for the over-constrained truck dock assignment problem with time windows and capacity constraint which attempts to assign trucks within its time window to docks to minimize shipping distances between docks. We have capacity constraint that the total number of pallets inside the crossdock cannot exceed the capacity of crossdock. Also note that in real world, cargo containers are huge in size, and one pallet usually carries exactly one cargo container at one time. Therefore, terms 'pallet', 'cargo', and 'cargo container' refer to the same transportation unit in the transshipment network

throughout the paper. The following notations are used:

- N : set of trucks arriving at and/or departing from the crossdock;
- M : set of docks available in the crossdock;
- n : total number of trucks, that is $|N|$, where $|N|$ denotes the cardinality of N ;
- m : total number of docks, that is $|M|$;
- a_i : arrival time of truck i ($1 \leq i \leq n$);
- d_i : departure time of truck i ($1 \leq i \leq n$);
- $w_{k,l}$: shipping distance for pallets from dock k to dock l ($1 \leq k, l \leq m$);
- $f_{i,j}$: number of pallets transferring from truck i to truck j ($1 \leq i, j \leq n$);
- C : capacity of crossdock, i.e. the maximum number of pallets the crossdock can hold at a time.

In addition, we use another dock, dock $m + 1$, which is rent from others for temporary usage when there is not enough capacity left inside the crossdock, or when all the docks are occupied. New arriving truck should go to this dock to load and unload cargoes, but it will incur a much higher cost (distance is longer). This dock can be used by many trucks simultaneously with infinite capacity. We regard it as a reasonable remedy to make the problem always feasible. Let $w_{k,m+1}$ ($w_{m+1,k}$) be shipping distance from dock k ($m + 1$) to dock $m + 1$ (k) ($1 \leq k \leq m$). Figure 1 illustrates an outline of the crossdock and major elements of our problem.

The following auxiliary variables are used:

$x_{i,j} \in \{0, 1\}$: 1 iff truck i departs no later than truck j arrives; 0 otherwise.

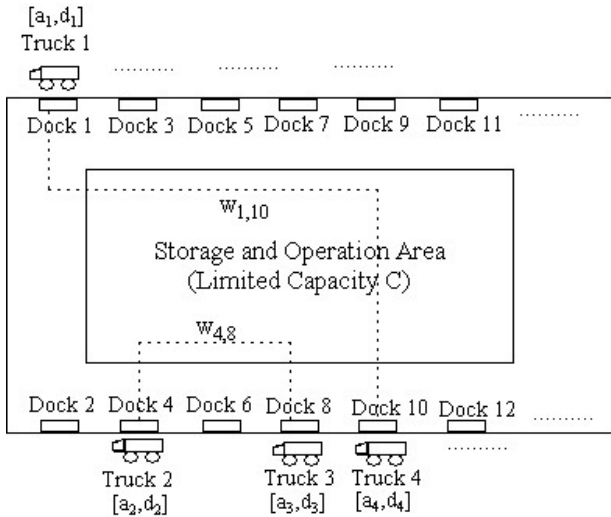


Fig. 1. Truck Dock Assignment Problem with Operation Time Constraint

The decision variables are as follows:

$y_{i,k} \in \{0,1\}$: 1 if truck i is assigned to dock k , 0 otherwise. ($1 \leq i \leq n$, $1 \leq k \leq m + 1$). $y_{i,k} = y_{j,k} = 1$ implies that $a_i > d_j$ or $a_j > d_i$ ($1 \leq i, j \leq n$);

$z_{ijkl} \in \{0,1\}$: 1 iff truck i is assigned to dock k and truck j is assigned to dock l ($1 \leq i, j \leq n, 1 \leq k, l \leq m + 1$).

Before we give out the model, in order to make the problem and data reasonable, some remarks of are made as follows:

- 1) $f_{i,j} \geq 0$, iff $d_j \geq a_i$ ($1 \leq i, j \leq n$), otherwise $f_{i,j} = 0$ which means truck i will transfer some cargo to truck j iff truck j departs no earlier than truck i arrives;
- 2) $a_i < d_i$ ($1 \leq i \leq n$) which means for each truck, the arrival time should strictly smaller than its departure time;
- 3) $n > m$ which satisfies the over-constrained condition;
- 4) capacity C is defined as follows: when truck i comes, it consumes units of capacity equal to $\sum_{k=1}^m \sum_{l=1}^m \sum_{j=1}^n f_{i,j} y_{i,k} y_{j,l}$. On the other hand, when truck j leaves, $\sum_{k=1}^m \sum_{l=1}^m \sum_{i=1}^n f_{i,j} y_{i,k} y_{j,l}$ units of capacity are released.
- 5) sort all the a_i 's and b_i 's in an increasing order, and let t_r ($r = 1, 2, \dots, 2n$) correspond to these $2n$ numbers such that $t_1 \leq t_2 \leq \dots \leq t_{2n}$. Use this notation, we can easily formulate the set of capacity constraints.

Our objective is to minimize the total shipping distance of transferring cargo between docks inside the crossdock. The IP model is as follows:

$$\min \sum_{k=1}^{m+1} \sum_{l=1}^{m+1} \sum_{i=1}^n \sum_{j=1}^n f_{i,j} w_{k,l} z_{ijkl}$$

s.t.

$$\sum_{k=1}^{m+1} y_{i,k} = 1 (1 \leq i \leq n) \tag{1}$$

$$z_{ijkl} \leq y_{i,k} (1 \leq i, j \leq n, 1 \leq k, l \leq m + 1) \tag{2}$$

$$z_{ijkl} \leq y_{j,l} (1 \leq i, j \leq n, 1 \leq k, l \leq m + 1) \tag{3}$$

$$y_{i,k} + y_{j,l} - 1 \leq z_{ijkl} (1 \leq i, j \leq n, 1 \leq k, l \leq m + 1) \tag{4}$$

$$x_{i,j} + x_{j,i} \geq z_{ijkk} (1 \leq i, j \leq n, i \neq j, 1 \leq k \leq m) \tag{5}$$

$$\sum_{k=1}^m \sum_{l=1}^m \sum_{i \in \{i: a_i \leq t_r\}} \sum_{j=1}^n f_{i,j} z_{ijkl} - \sum_{k=1}^m \sum_{l=1}^m \sum_{i=1}^m \sum_{j \in \{j: d_j \leq t_r\}} f_{i,j} z_{ijkl} \leq C (1 \leq r \leq 2n) \tag{6}$$

$$y_{i,k} \in \{0, 1\}, y_{j,l} \in \{0, 1\}, z_{ijkl} \in \{0, 1\} (1 \leq i, j \leq n, 1 \leq k, l \leq m + 1) \quad (7)$$

Constraints (1) ensures that each truck must be assigned to exactly one dock. Constraints (2)-(4) jointly define the variable z which represent the logic relationship among $y_{i,k}$, $y_{j,l}$ and z_{ijkl} . Constraint (5) specifies that one dock cannot be occupied by two different trucks simultaneously. Finally, constraint (6) is capacity constraint which means that for each time point t_r , the total number of pallets inside the crossdock cannot exceed the capacity C .

3 Tabu Search

Tabu search (TS) is a heuristic search procedure that proceeds iteratively from one solution to another by moves in a neighborhood space with the assistance of adaptive memory [4]. We next describe TS memory and the framework used for the problem.

3.1 TS Memory

The solution is represented by a sequences A , which has length n (n is the number of trucks). The sequence A represents the dock assignment. For example, consider an instance with m docks and n trucks. The solution is then a sequence (s_1, s_2, \dots, s_n) , which means that Truck 1 is assigned to dock (gate) s_1 , Truck 2 is assigned to dock s_2, \dots , Truck n is assigned to dock s_n ($0 \leq s_i \leq m, 1 \leq i \leq n$). If the truck i is unassigned to any of the docks, which is possible when all the docks are occupied, we give s_i the value of 0. If the solution is feasible, the dock assignment is then uniquely determined by the sequence of (s_1, s_2, \dots, s_n) . The representation is depicted in Fig. 2. In the TS memory we implement, only the assignment information is captured so that only the move that has the identical neighborhood exchange move to the assignments will be forbidden.

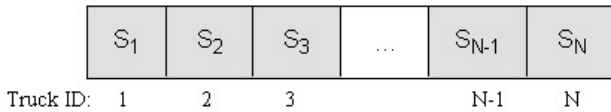


Fig. 2. The Solution Representation

3.2 Neighborhood Search

We use the a modified neighborhood search approach from Ding [3], which consists of three moves:

The Insert Move: Move a single truck to a dock gate other than the one it currently assigns. It is depicted in Fig. 3.

The Interval Exchange Move: Exchange two truck intervals in the current assignment. A truck interval is a group of consecutive trucks assigned to one dock gate. The move is depicted in Fig. 4.

The Rented Dock Exchange Move: Exchange a truck which is assigned to the rented dock with a truck that is assigned to a general dock gate.



Fig. 3. The Insert Move

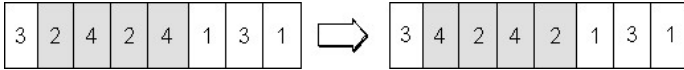


Fig. 4. The Interval Exchange Move

3.3 TS Framework

The TS algorithm can be described by the following steps:

1. Generate an initial solution x_{init} randomly, set $x_{curr} \leftarrow x_{init}$;
2. Generate a set of neighborhood solutions $N(x_{curr})$ of x_{curr} by the Insert Move and Interval Exchange Move;
3. The solution $x \in N(x_{curr})$ with the least cost and satisfying either one of the two conditions (1) and (2) and must satisfying condition (3) will be selected: (1) it is not forbidden (i.e. the assignment is not identical to any assignments of recent tabu tenure moves); (2) The cost of x is better than the current best cost (aspiration criterion); (3) The occupied capacity of x at boundary time points of all time windows is less than the capacity of crossdock C .
4. Set $x_{curr} \leftarrow x$; update the TS memory;
5. If the termination conditions are satisfied, stop; otherwise jump to step 2;

When we generate the neighborhood solutions, we randomly choose three types of moves with equal probability. There are two termination conditions: either the best solution cannot be improved within a certain number of iterations, or the maximum number of iterations has been reached.

4 Genetic Algorithm

Genetic algorithms (GA) have become a well-known meta-heuristic approach for difficult combinatorial optimization problems [5]. In this second approach to the dock assignment problem, we found it suitable to solve it. We first discuss some essential components of GA, including solution representation and crossover operators, and then outline the framework of our GA.

4.1 Solution Representation

The chromosome is an important component in GA and has great influence on the algorithm output. In the basic GA, a chromosome is usually encoded as a sequence and represents a solution. Similar to the TS solution representation (see Fig. 2), we here represent the dock assignment solution by a chromosome sequence that defines an assignment of trucks to the docks. During the population

initialization process, if the randomly generated chromosome is infeasible, we just drop the infeasible sequence and generate a new one.

4.2 Crossover and Mutation

In our problem, chromosomes are not permutation sequences such as in the Travelling Salesman Problem. Hence, well-known crossover operators, such as Partially Mapped Crossover and Cycle Crossover, cannot be used. We implemented two crossover operators: One-Point Crossover and Two-Point Crossover.

In the One-Point Crossover, one random crossover point is selected. The first part of the first parent is attached with the second part of the second parent to make the first offspring. The second offspring is built from the first part of the second parent and the second part of the first parent. The following is an example of One-Point Crossover operator (the crossover point is denoted by |):

Parent1: (3 2 4 2 4 | 1 3 1), Parent2: (2 1 2 4 3 | 2 2 4)
 Offspring1: (3 2 4 2 4 | 2 2 4), Offspring2: (2 1 2 4 3 | 1 3 1)

In the Two-Point Crossover, two random crossover points are selected for one crossover operation. The chromosomal materials are swapped between two cut points to produce offsprings. This is illustrated in the following example:

Parent1: (3 2 | 4 2 4 | 1 3 1), Parent2: (2 1 | 2 4 3 | 2 2 4)
 Offspring1: (3 2 | 2 4 3 | 1 3 1), Offspring2: (2 1 | 4 2 4 | 2 2 4)

In the proposed GA, we use both of the above-mentioned cross over operators with equal probability. We chose the ‘Swap Mutation’ as our mutation operator, which selects two positions at random and swaps the values at those positions. For example, the following mutation swaps the values at position 3 and position 6: (3 2 4 2 4 1 3 1) \rightarrow (3 2 1 2 4 4 3 1). For simplicity, we do not apply any repair function to the infeasible offsprings. Therefore, if an offspring is infeasible by violating constraints in the problem, we simply remove it from the GA population base.

4.3 GA Framework

With these components of GA, we now outline GA as follows. In this algorithm, #pop, #crossover, #iter and p_1 are parameters which are specified within experiments.

```

Initialize Pop with size #pop
for iter  $\leftarrow$  1 to #iter do
  for off  $\leftarrow$  1 to #crossover do
    Randomly select ParentA and ParentB
    Crossover ParentA and ParentB to produce OffspringA and OffspringB
  end for
  for each new-produced individual indv do
    mutate indv with probability  $p_1$ 
  end for

```

```

evaluate each individual
select the best  $\#pop$  individuals from all the individuals
update current best solution
end for
    
```

5 Experimental Results and Analysis

All the algorithms were coded in Java and tested on a P4 2.4GHz PC with 512M RAM. As comparison, we use ILOG CPLEX 8.0 to the IP formulation presented in Section 2. The test generation process, parameter settings of various methods, and detailed computational results are presented in the following subsections.

5.1 Test Data and Experiment Setup

We chose a representative layout of a crossdock to have two parallel sets of terminals, where docks are symmetrically located in the two terminals shown in Fig. 5. We set the distance between two adjacent docks within one terminal (e.g., dock 1 and dock 3) to be 1 unit and the distance between two parallel docks in different terminals (e.g., dock 1 and dock 2) to be 3 units. To simplify the problem, we assumed that forklift can only walk ‘horizontally’ or ‘vertically’, i.e., if one forklift wants to transfer one pallet from dock 3 to dock 2, his walking distance is $1+3=4$ units. This is similar to the so-called Manhattan metric. In addition, we set the distance between rented dock and any of docks inside the crossdock to be 10 in order to make it undesirable to use.

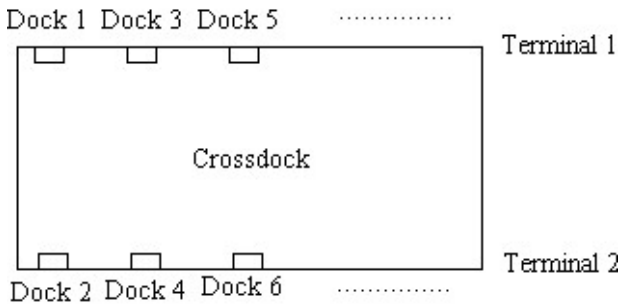


Fig. 5. Crossdock topology

The start points of truck time window $a_i (1 \leq i \leq n)$ are uniformly generated in the interval $[1, \frac{n*70}{m}]$. The end points b_i are generated as $b_i = a_i + [45, 74]$. The number of transferring pallets $f_{i,j}$ is randomly generated in the interval $[6, 60]$ if $d_j \geq a_i$ (0 otherwise). In TS, Maximum number of iterations is 10^6 and each time 100 neighbors are generated with a tabu tenure = 10. The algorithm is to terminate if the best solution was not improved within 10^4 iterations. In GA, we

specify $\#iter = 10^4$, $\#pop = 300$, and $\#crossover = 500$. The mutation probability p_1 is taken to be 0.2. The maximum iteration is 10^5 and the termination condition was when the best solution did not improve within 500 iterations.

5.2 The Results

We designed two categories of test sets. The detailed results are presented in Table 1–2 respectively. The first row of each table denotes the instance ID or group ID. The second row contains the sizes, where $n \times m$ means that there are n trucks and m docks. The rest of the rows provide the results of various methods proposed in this paper. Each result cell contains two values. The value on the top provides the result, whereas the value at the bottom provides the computational time in seconds.

1. Small-scale Test Sets

Small-scale test sets are generated with the size ($n \times m$) ranging from 12×4 to 18×7 . We see that GA performs extremely well for this group of test sets, as it obtains the best objective values for all test sets. Solution quality of TS are slightly worse than GA while the runtime of TS is considerably faster. At last, CPLEX gets the optimal solutions only for the three smallest test sets within the time limit of 2 hours. For all other 7 test sets that CPLEX exceeds the time limit, the solution quality is generally worse than the two meta-heuristic approaches.

2. Large-scale Test Sets

Large-scale test sets ranging from 20×6 to 35×9 are used here. We see from Table 2. that the performance of TS surpasses GA in both solution quality and runtime, but in general, both meta-heuristics have provided quite good solutions. beating GA in 8 of the 10 test sets. We also note the solution time of GA grows very quickly as the problem size expands, which indicates that GA does not scale quite well for the problem.

As a whole, we can conclude from the experiment that both two proposed meta-heuristics are effective approaches for the dock assignment problem. If a minimal runtime of solving medium-to-large-scale instances is required, TS is then more preferred than GA.

Table 1. Results: Small-scale Test Sets

Size	12 X 4	12 X 5	14 X 4	14 X 5	16 X 6	16 X 7	18 X 6	18 X 7
CPlex	<u>7523</u>	<u>8914</u>	<u>12117</u>	3680	22984	13049	14712	14276
Time(s)	1035	5460	3166	≥ 7200	≥ 7200	≥ 7200	≥ 7200	≥ 7200
TS	<u>7523</u>	8917	12200	3578	21017	12528	13760	13818
Time(s)	0.55	0.58	0.56	0.58	0.66	0.64	0.72	0.78
GA	<u>7523</u>	<u>8914</u>	<u>12117</u>	<u>3464</u>	<u>20953</u>	<u>12489</u>	<u>13635</u>	<u>12986</u>
Time(s)	1.89	1.75	2.53	2.61	2.91	4.13	4.81	5.58

Table 2. Results: Large-scale Test Sets

Size	20 X 6	20 X 7	25 X 6	25 X 7	30 X 8	30 X 9	35 X 8	35 X 9
CPlex	62537	64682	101683	112648	146595	134762	148654	145725
Time(s)	≥ 7200	≥ 7200	≥ 7200	≥ 7200	≥ 7200	≥ 7200	≥ 7200	≥ 7200
TS	<u>40777</u>	<u>49219</u>	<u>68610</u>	<u>76418</u>	99442	<u>97981</u>	<u>105312</u>	116147
Time(s)	1.06	1.02	1.31	1.28	1.84	1.81	2.28	2.34
GA	42405	60610	68892	78112	<u>95918</u>	98355	105768	<u>111341</u>
Time(s)	5.36	1.95	9.95	7.20	17.83	15.80	51.11	59.00

6 Conclusions and Future Work

In this paper, we consider the over-constrained truck dock assignment problem with time windows and capacity constraint in transshipment network through crossdocks where the number of trucks exceed the number of docks available and the capacity of the crossdock is limited, and where the objectives are to minimize the total shipping distances. The problem is formulated as an IP model. We then propose a Tabu Search (TS) and a Genetic algorithms (GA) that utilize the IP constraints. Experiments were conducted using a range of test data sets that reflect realistic scenarios. The heuristic search algorithms are compared with the CPLEX solver, showing they obtain better results within shorter running times. In the future, we think that although this problem considers crossdock optimization problem, it can also be applied to the air gate assignment problem in airport to schedule the flights well in order to achieve better performance. The other direction of research is to study a realistic model which take the operational time, instead of the distance, of each pallet into consideration. Therefore, the arrival/departure time windows of the trucks can be well related to the inner-crossdock operations.

References

- [1] J.E. Aronson. A survey on dynamic network flows. *Annals of Operations Research*, 20:1–66, 1989.
- [2] J. Bartholdi and K. Gue. Reducing labor costs in an ltl cross-docking terminal. *Operations Research*, 48:823–832, 2000.
- [3] H. Ding, A. Lim, B. Rodrigues, and Y. Zhu. The over-constrained airport gate assignment problem. *Computers and Operational Research*, 32:1867–1880, 2005.
- [4] F. Glover and M. Laguna. *Tabu Search*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.
- [5] John H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI, 1975.
- [6] A. Lim, B. Rodrigues, and Y. Zhu. Airport gate scheduling with time windows. *Artificial Intelligence Review*, 24:5–31, 2005.
- [7] T. Obata. The quadratic assignment problem: Evaluation of exact and heuristic algorithms. Technical report, 2000.
- [8] L. Tsui and C. Chang. Optimal solution to dock door assignment problem. *Computers and Industrial Engineering*, 23:283–286, 1992.

An Entropy Based Group Setup Strategy for PCB Assembly

In-Jae Jeong

Department of Industrial Engineering, Hanyang University,
Seoul, 133-791, South Korea
ijeong@hanyang.ac.kr

Abstract. Group setup strategy exploits the PCB similarity in forming the families of boards to minimize makespan that is composed of two attributes, the setup time and the placement time. The component similarity of boards in families reduces the setup time between families meanwhile, the geometric similarity reduces the placement time of boards within families. Current group setup strategy considers the component similarity and the geometric similarity by giving equal weights or by considering each similarity sequentially. In this paper, we propose an improved group setup strategy which combines component similarity and geometric similarity simultaneously. The entropy method is used to determine the weight of each similarity by capturing the importance of each similarity in different production environments. Test results show that the entropy based group setup strategy outperforms existing group setup strategies.

Keywords: Printed circuit board assembly, group setup, entropy method, similarity coefficient.

1 Introduction

This paper considers a group setup problem in a single SMT machine producing multiple types of boards. The head starts from a given home position, moves to feeder carriage on the machine to pick up the component. After picking up the component, the head moves to the placement location on the PCB for this component. Then the component is placed on the board and the head travel back to the feeder carriage to pick up the next component. The pick-and-place process continues until all components required for the board have been completed.

Let K be the total number of family and N_f be the number of boards in family f . Then the total number of boards, $N = \sum_{f=1}^K N_f$. We assume that the head velocity, v (mm/sec) and the feeder installation/removal time, σ are constant for all types of boards. Also, let m_f be the number of feeder changes required from family $f - 1$ to f and d_i be the length of tour followed by the head to assemble board i . b_i is the batch size of board i . Leon and Peters (1996) proposed the following conceptual formulation of the group setup problem:

Minimize: $\text{Makespan} = \sum_{f=1}^K (\sigma m_f + \sum_{i=1}^{N_f} \frac{b_i}{v} d_i)$
 Subject to: Feeder capacity constraints
 Component-feeder constraints
 Component placement constraints

The objective is to minimize the makespan for producing multiple types of boards. The first term of the makespan is the setup time to remove the previous setups and install components on feeders for current family. The second term is the time to place all components on all boards in a batch for current family. If all boards are grouped as a single family, the setup will occur only once minimizing setup time. However, the single family solution will increase the total placement time since the common setup is not prepared for individual boards. On the other hand, if all boards form a unique family of its own, the placement time reduction will be surpassed by setup time. Hence, boards must be grouped such that within the family, boards share as many common component types as possible (i.e., component similarity) in order to reduce setup time between families. Also the placement locations of boards within the family must be similar to each others (i.e., geometric similarity) in order to reduce placement time. Therefore the development of good similarity coefficient is important issue in a group setup strategy.

The decision variables are the number of family K , the types of boards in family f , N_f and the placement sequence of locations in board i and the component-feeder assignment for family f to determine d_i .

The first constraints represent the feeder capacity constraints. Total number of different component types in any family can not exceed the feeder capacity since only one component type can reside in one feeder slot. The second constraints, component-feeder constraints means that each component needed for boards in a family must be assigned to a feeder. The third constraints, component placement constraints are equivalent to traveling salesman problem (TSP) constraints. That is, the placement head must visit all the placement locations on a board. The distance between two placement locations is the time for the head to move from the first placement location to the feeder slot containing component for the second placement then to the second placement location.

Existing group setup strategies (1) considers component similarity only (Leon and Peter 1998) or (2) forms families of boards based on geometric similarity and select the groups of boards based on component similarity in sequential manner (Leon and Jeong 2005) or (3) considers an overall board's similarity coefficient which combines component similarity and geometric measure by assigning equal weights (Quintana and Leon 1999). Leon and Jeong (2005) reported that the performance of group setup strategy of case (2) performs better than other cases.

The motivation of this paper was the belief that the determining appropriate weights of case (3) and combining both similarities simultaneously could achieve a further reduction of makespan . Combining different criteria into a synthesized criterion falls into a well known research area, Multiple Criteria Decision Making (MCDM). In this paper, we use the entropy method for calibrating the weights assigned to the component similarity and the geometric similarity. The entropy

concept suggests that if the component similarity or the geometric similarity of boards is the same, the similarity can be eliminated from further considerations in forming the families of boards. Alternately, the weight assigned to a similarity is small if all boards have the similar value of corresponding similarity coefficient.

2 Backgrounds

There are a number of different PCB setup strategies to reduce makespan in the literature (i.e., unique setup, minimum setup, group setup, partial setup). In this section, we focus on partial setup and group setup because the partial setup performs better than other setup strategies (Leon and Peters, 1996) and the implementation of group setup is relatively easier than partial setup in real world. The procedure for partial setup strategy (Leon and Peters, 1996) is summarized in the following steps.

Partial setup procedure

Step 1: Determine an arbitrary board sequence

Step 2: Repeat for a given number of times

Step 3: For each board.

Step 4: Find a feasible component-feeder assignment

Step 5: Repeat for a given number of times

Step 6: Find a placement sequence given a component-feeder assignment determined at Step 4

Step 7: Find a component-feeder assignment given a placement sequence determined at Step 6

Step 8: Determine the matrix of sequence-dependent changeover times. Sequence-dependent setup time is the time it takes to remove and install necessary feeders when changing from a board to another board.

Step 9: Determine the board sequences that minimizes the total changeover time given the sequence dependent changeover time.

As shown in Step 8, portions of the previous setup may remain intact when changing over between boards in partial setup. Therefore only a portion of components are removed and installed between boards which might be a complicated operation. However, in group setup, once all the family has been assembled, all of the components are completely removed from feeder slots. The traditional group setup procedure is summarized in the following steps (Leon and Peters, 1996).

Group setup procedure.

Phase 1: Clustering (Form K families of boards with similar boards. Family sizes can not exceed the maximum number of feeder slots.)

- Step 1: Put each board-type in a single-member family
 Step 2: Compute similarity coefficient, s_{ij} for all pairs of family i and j
 Step 3: Compute clustering objective values
 Step 4: Set $T = \max(s_{ij})$
 Step 5: Merge the pair of board i^* and j^* , if $s_{i j} = T$. Repeat until no more pairs can be merged at similarity level T .
 Step 6: Compute clustering objective and save the clustering solution if an improvement was achieved.
 Step 7: Repeat Step 2 through 6 while merging is possible.

Phase 2: Component-feeder assignment and placement sequence.

- Step 8: Form a composite-board $H(f)$, $f=1, \dots, K$, this board consists of the superposition of all the placement locations with their corresponding components of the boards in family f .
 Step 9: Determine a feasible component-feeder assignment $C(H_f)$
 Step 10: For all $i \in N_f$, find a placement sequence $P(i)$, given $C(H_f)$
 Step 11: For all $i \in N_f$, find a component-feeder assignment $C(H_f)$ given $P(i)$
 Step 12: Repeat Step 10 and Step 11 for a predetermined number of iterations.

In Phase 1, the hierarchical clustering algorithm merges similar boards into a family. The clustering procedure continues until all boards form a single family. To form good families of boards, it is essential to develop a similarity coefficient which considers both the component similarity and the geometric similarity of any two boards. Another issue in hierarchical clustering is the development of clustering objective in order to evaluate the quality of board clustering (e.g., minimization of the similarity coefficient between families, maximization of the similarity coefficient within families).

In phase 2, we consider each family as a single composite-board, H_f and determine the component-feeder assignment and placement sequence. For a given component-feeder assignment, $C(H_f)$, the placement sequencing problem can be solved as TSP problems. In this paper, we use the nearest-neighbor heuristic to solve the TSP. For a given placement sequences, $P(i)$, the component-feeder assignment problem is a LAP. In this implementation, the LAP is solved using the shortest augmenting path algorithm proposed by Jonker and Vogenant (1987). The LAP/TSP heuristic terminates when it reaches the predetermined number of iteration.

Currently, there exists two group setup strategies (i.e., Placement Location Matrix (PLM) based group setup strategy (Quintana and Leon, 1998) and Minimum Metamorphic Distance (MMD) based group setup strategy (Leon and Jeong, 2005)) which consider both component similarity and geometric similarity in the literature. Each strategy uses the same frame work of two phase procedure except the definition of the similarity coefficient in Step 2 and the clustering objective in Step 5. PLM based group setup strategy uses the following board's similarity coefficient.

s_{ij} : similarity of board i and j.

$x^{i \cap j}$: Number of Common Component (NCC) types between board i and j.

$x^{i \cup j}$: total number of different component types required by board i and j.

D_{ij} : dissimilarity of board i and j.

F_{ij} : frequency ratio of the number of placement locations between board i and j.

$\sqrt{X^2 + Y^2}$:point magnitude of coordinate (X,Y).

p_{ki} : point magnitude of k th sorted placement location in ascending order for board i .

n_i : number of placement location of board i .

NP_j :number of placement locations of board j .

$n^* = \min(n_i, n_j)$.

Xrange, Yrange: Cartesian distance of the largest board.

$$s_{ij} = 0.5s_{ij}^{NCC} + 0.5(1 - D_{ij})F_{ij} \tag{1}$$

where $s_{ij}^{NCC} = \frac{x^{i \cap j}}{x^{i \cup j}}$, $D_{ij} = \frac{\sqrt{\sum_k^n (p_{ki} - p_{kj})^2}}{n \sqrt{Xrange^2 + Yrange^2}}$, $F_{ij} = \frac{\min(NP_i, NP_j)}{\max(NP_i, NP_j)}$,

The nominator of D_{ij} measures the dissimilarity of the magnitude of boards and the denominator is the normalizing factor. Therefore $(1 - D_{ij})$ represents the similarity measure of board i and j. F_{ij} measures the frequency ratio of the number of placement locations between two boards. Therefore, two boards with the same number of placement locations are strongly associated.

There are some limitations on PLM methods. First, point magnitude of two different points could be the same. For example, point magnitude of (a,b) is the same as the one of (b,a). This could be wrongly interpreted such that there is no dissimilarity between two points. Secondly, giving equal weights for similarities may not appropriate in cases where the placement time becomes more important than the setup time or vice versa in reducing makespan.

A sequential treatment of component similarity and geometric similarity has been proposed by Leon and Jeong (2005) namely, Minimum Metamorphic Distance (MMD) based group setup strategy. Suppose that board i and j have the same number of placement locations of component type c . Then the Euclidean distance matrix from locations in board i to board j can be constructed. The problem is to find the best assignment of from-to locations which minimize the total sum of Euclidean distance (MMD_{ij}^c). The solution can be easily found using LAP method. When boards with different number of locations are used, all the locations on the board with more locations are assigned to the locations on the board with less number of locations. In MMD based setup, a new geometric similarity has been proposed as follows:

MMD_{ij}^c : minimum metamorphic distance of board i and board j for component type c .

p : placement locations of board i.

q :placement locations of board j.

d_{pq}^c : Euclidean distance between location p and q with component type c .

$$s_{ij}^{MMD} = 1 - \frac{\sum_{\forall c} MMD_{ij}^c}{\sum_{\forall c} \sum_{\forall p} \max_{\forall q} (d_{pq}^c)} \tag{2}$$

As shown in equation (2), when MMD increases, the geometric similarity decreases. The authors suggested a group setup strategy considering the component similarity (i.e., s_{ij}^{NCC} in equation (1)) and the MMD based geometric similarity (i.e., s_{ij}^{MMD} in equation (2)) sequentially. In hierarchical clustering, the proposed procedure merges two boards with the largest MMD similarity. Then the clustering objective is the maximization of average s_{ij}^{MMD} within families per unit feeder change between families. Therefore, the clustering objective is maximized when all boards in families are geometrically similar (i.e., placement time is minimized) and the number of feeder change is minimized (i.e., setup time is minimized). The limitation of the MMD based group setup is that the component similarity and the geometric similarity are not considered simultaneously. Forming the families of boards considering only geometric similarity may reduce the possibility of generating solutions which is favorable in reducing setup time. However the authors reported that MMD based group setup outperformed the PLM based group setup. In section 3, we propose a new group setup strategy which combines s_{ij}^{NCC} and s_{ij}^{MMD} using the entropy method.

3 Entropy Based Group Setup Strategy

In the past two decades, there has been of enormous growth in the area of multi-attributes optimization. One of the most important issue in this research area is the development of appropriate weights for different attributes. As each attribute has different scale, synthesizing attributes by giving appropriate weights to each attribute is essential to solve the optimization problem. The entropy method suggests that the weight assigned to a criterion must be small if all alternatives have similar value for the criterion. On the other hand, when the difference between a criterion’s values is great, the criterion must be considered as important by giving large weight. Let

$NCC_{ij} = x^{i \cup j}$: number of common component type between board i and board j.

$MMD_{ij} = \sum_{\forall c} MMD_{ij}^c$: minimum metamorphic distance between board i and board j $\forall i, \forall j, i \neq j$.

Then the entropy measures of the criteria for NCC and MMD are as follows:

$$e(NCC) = - \sum_{i=1}^N \sum_{j=1}^N \frac{NCC_{ij}}{S_{NCC}} \ln \frac{NCC_{ij}}{S_{NCC}} \tag{3}$$

$$e(MMD) = - \sum_{i=1}^N \sum_{j=1}^N \frac{MMD_{ij}}{S_{MMD}} \ln \frac{MMD_{ij}}{S_{MMD}} \tag{4}$$

where $S_{NCC} = \sum_{i=1}^N \sum_{j=1}^N NCC_{ij}$, $S_{MMD} = \sum_{i=1}^N \sum_{j=1}^N MMD_{ij}$. When all NCC_{ij} are equal, then $\frac{NCC_{ij}}{S_{NCC}} = \frac{2}{N(N-1)}$ and the maximum of $e(NCC)$ is achieved which is $e_{max}(NCC) = \ln \frac{N(N-1)}{2}$. This implies that if the value of a criterion is evenly distributed, then the entropy of the criterion is maximized and the entropy is minimized when the criterion value is biased. By setting a normalization factor, $K = \frac{1}{e_{max}(NCC)} = \frac{1}{\ln(\frac{N(N-1)}{2})}$, $0 \leq e(NCC) \leq 1$ can be achieved. Therefore the normalized entropy measures of equation (3) and (4) are

$$e(NCC) = -K \sum_{i=1}^N \sum_{j=1}^N \frac{NCC_{ij}}{S_{NCC}} \ln \frac{NCC_{ij}}{S_{NCC}} \tag{5}$$

$$e(MMD) = -K \sum_{i=1}^N \sum_{j=1}^N \frac{MMD_{ij}}{S_{MMD}} \ln \frac{MMD_{ij}}{S_{MMD}} \tag{6}$$

We impose a large weight for a criterion when the corresponding entropy measure is small since the information transmitted by the criterion is great (i.e., there exists great difference between the values of the criterion). The weights are calculated as follows:

$$W_{NCC} = \frac{1 - e(NCC)}{2 - (e(NCC) + e(MMD))} \tag{7}$$

$$W_{MMD} = \frac{1 - e(MMD)}{2 - (e(NCC) + e(MMD))} \tag{8}$$

Using the entropy method, we propose a board’s similarity coefficient of board i and j as follows;

$$s_{ij} = W_{NCC} s_{ij}^{NCC} + W_{MMD} s_{ij}^{MMD} \tag{9}$$

where s_{ij}^{NCC} is the component similarity as shown in equation (1) and s_{ij}^{MMD} is the MMD based geometric similarity as shown in equation (2). It is important to note that the entropy method can easily be extended to the development of board’s similarity coefficient with more than two criteria.

The entropy based group setup strategy uses the generic group setup procedure in section 2 with the board’s similarity coefficient in equation (9). The clustering objective is as the same as the one of MMD based setup (i.e., maximization of average similarity within families per unit number of feeder change between families).

4 Experiments

In this paper, we consider a generic machine that has 70 feeder slots with 20mm between the slots. The board dimensions are maximum 320mm 245mm and the coordinates for each board were randomly generated from uniform distributions as follows: $X=635+U(0,245)$, $Y=254+U(0,320)$. The home position coordinate is

(0,0) and the first feeder slot location is (457,0). The number of component types required per board were generated from $U(6,20)$ from 70 different component types. We considered the time to install or remove feeder, in cases of 30(sec) and 60(sec). The head velocity, v was tested for 100(mm/sec) and 300(mm/sec). The batch size of boards, b were generated from $U[50,100]$. Also the total number of boards, N were generated from $U[5,15]$. The placement locations and the corresponding component types were generated from a seed board. A seed board is created with location $(L_{sx}(i), L_{sy}(i))$ where $L_{sx}(i)$ is the x-coordinate of i th placement location for seed board and $L_{sy}(i)$ is the y-coordinate. $C_s(i)$ is the component types of i th placement location for the seed board. We fixed the number of placement location to 50 for the seed board. Based on the component similarity (C) and geometric similarity (G), another board (i.e., a child board) is created using the following formula;

$$L_{cx}(i) = L_{sx}(i) + (1 - G) \times 0.5 \times 245 \times U(-1, 1) \tag{10}$$

$$L_{cy}(i) = L_{sy}(i) + (1 - G) \times 0.5 \times 320 \times U(-1, 1) \tag{11}$$

$$C_c(i) = \begin{cases} C_s(i) & \text{with probability } C \\ U(1, NC_c) & \text{otherwise} \end{cases} \tag{12}$$

Where $L_{cx}(i)$ is the x-coordinate of i th placement location for child board and $L_{cy}(i)$ is the y-coordinate. $C_c(i)$ is the component types of i th placement location for the child board. NC_c is the number of component type of the child board c . Based on these experimental factors and parameters, we generated 16 problem types as shown in Table 4. Each problem set consists of 20 random problems.

Table 1. Problem types

Problem type	Head velocity (mm/sec)	Feeder change time (sec)	Component similarity (C)	Geometric similarity (G)	Problem type	Head velocity (mm/sec)	Feeder change time (sec)	Component similarity (C)	Geometric similarity (G)
1	100	30	0.2	0.75	9	100	30	0.2	0.2
2	100	30	0.75	0.75	10	100	30	0.75	0.2
3	100	60	0.2	0.75	11	100	60	0.2	0.2
4	100	60	0.75	0.75	12	100	60	0.75	0.2
5	300	30	0.2	0.75	13	300	30	0.2	0.2
6	300	30	0.75	0.75	14	300	30	0.75	0.2
7	300	60	0.2	0.75	15	300	60	0.2	0.2
8	300	60	0.75	0.75	16	300	60	0.75	0.2

To measure the performance of the different setup strategies, the deviation from partial setup is computed as follows:

$$\text{Percent deviation from partial setup} = \frac{M^{\text{setup strategy}} - M^{PS}}{M^{PS}} \times 100\%$$

Percent M^{PS} represents the makespan of partial setup (PS) strategy. $M^{\text{setup strategy}}$ corresponds to the makespan of PLM based group setup (PLM), MMD based group setup (MMD) and the Entropy based group setup (ENT).

Table 2 summarizes the average setup time, average placement time and average makespan of different setup strategies. Consider problem type 1 where head velocity is 100mm/sec, feeder change time is 30sec, component similarity is 20% and geometric similarity is 75%. Note that in this specific problem, the placement time is more important than setup time since the head velocity is slow, the feeder change time is short. Result shows that PLM performs better in terms of setup time than PS, MMD and ENT. This is because PS, MMD and ENT achieve an improvement in the reduction of the placement time instead of setup time. In addition ENT assigns the larger weight for the geometric similarity of boards than MMD under consideration. As a result, ENT dominates PLM and MMD by reducing about 8% and 4% of makespan relatively as shown in Table 2. In summary, test results show that ENT outperforms PLM and MMD in terms of makespan. The maximum percent deviation from PS of ENT is 2.72% while MMD and PLM are 5.6% and 9.35% respectively. This result implies that ENT balances the tradeoff between the setup time and the placement time and finds the solution that minimizes the makespan for all types of problems.

Table 2. Summary of experimental results

Problem type	(PLM+PS)/PS*100'			(MMD+PS)/PS*100'			(ENT+PS)/PS*100		
	Setup time Average	Placement time Average	Makespan Average	Setup time Average	Placement time Average	Makespan Average	Setup time Average	Placement time Average	Makespan Average
1	-25.15	9.51	8.69	3.93	3.45	3.46	9.61	1.04	1.25
2	-21.58	1.21	0.86	-23.51	1.38	1.00	-10.64	0.90	0.72
3	-16.64	8.37	7.37	11.91	2.97	3.32	24.00	-0.30	0.66
4	-3.97	0.95	0.83	-10.99	1.52	1.21	-5.79	0.93	0.76
5	-15.67	8.61	7.25	12.25	4.89	5.30	34.44	-0.66	1.30
6	6.26	1.26	1.43	-3.66	0.97	0.81	15.17	0.53	1.01
7	14.70	0.19	1.02	6.94	0.61	1.09	18.67	0.24	1.29
8	21.71	-0.03	1.19	15.94	0.38	1.25	25.75	-0.11	1.35
9	-22.17	8.96	8.24	2.69	4.46	4.42	16.37	-0.04	0.34
10	-23.50	1.14	0.73	-26.87	1.66	1.19	-11.45	1.23	1.02
11	-24.37	10.79	9.35	4.97	4.93	4.94	13.45	2.26	2.72
12	-2.62	0.78	0.70	-8.26	1.25	1.01	10.06	0.72	0.96
13	-12.87	7.93	6.74	7.01	5.52	5.60	22.69	0.71	1.97
14	-5.38	1.36	1.12	0.67	0.86	0.85	12.31	0.22	0.65
15	12.55	0.04	0.75	12.99	0.10	0.83	39.29	-0.40	1.86
16	36.38	0.11	0.99	-21.54	1.44	0.89	-4.39	0.92	0.79
Overall									
Average	-5.15	3.82	3.58	-0.86	2.27	2.32	13.10	0.51	1.17
Min	-25.15	-0.03	0.70	-26.87	0.10	0.81	-11.45	-0.66	0.34
Max	36.38	10.79	9.35	15.94	5.52	5.60	39.29	2.26	2.72

*Results from Leon and Jeong (2005)

5 Conclusions

This paper has presented an improved group setup strategy based on entropy method considering both component similarity and geometric similarity. It has demonstrated how the entropy method determines weights for different criteria to adapt to a variety of production conditions. The improved group setup strategy dominated PLM or MMD based group strategy for all types of problems. Overall, improved group setup strategy deviated from partial setup, maximum 2.72% and average 1.17%. Future research includes the extension of the multiple SMT machines and the consideration of multiple criteria in grouping PCBs (e.g., due dates).

Acknowledgments

This research was supported by the research fund of Hanyang University (HY-2003). The author thanks Professor V. Jorge Leon at Texas A&M University for constructive suggestions.

References

1. Askin, R. G., Dror, M., and Vakharia, A. J.: Printed circuit board scheduling and component loading in a multimachine, openshop manufacturing cell. *Naval Research Logistics* **41**(5) (1994) 587–608
2. Ball, M. O. and Magazine, M. J.: Sequencing of insertions in printed circuit board assemblies. *Operations Research* **36**(2) (1988) 192–201
3. Chan, D. and Mercier, D. :IC Insertion: An Application of the Traveling Salesman Problem. *International Journal of Production Research* **27**(10) (1989) 1837–1841
4. Crama, Y.:Combinatorial optimization models for production scheduling in automated manufacturing systems. *European Journal of Operational Research* **99** (1997) 136–153
5. Crama, Y., Kolen, A. W. J., Oerlemans, A. G., and Spieksma, F. C. R.: Throughput rate optimization in the automated assembly of printed circuit boards. *Annals of Operations Research* **26** (1990) 455–480
6. Jonker, R. and Volgenant, A.: A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing* **59** (1987) 231–340
7. Leon, V. J. and Jeong I. J.:An improved group setup strategy for PCB assembly. *Lecture Notes in Computer Science*. **3483** (2005) 312–321
8. Leon, V. J. and Peters B. A.: Re-planning and analysis of partial setup strategies in printed circuit board assembly systems. *International Journal of Flexible Manufacturing Systems Special Issue in Electronics Manufacturing* **8**(4) 1996 389–412
9. Leon, V. J. and Peters B. A.: A comparison of setup strategies for printed circuit board assembly. *Computers in Industrial Engineering* **34**(1) (1998) 219–234
10. Lofgren, C. B. and McGinnis, L. F.: Dynamic scheduling for flexible printed circuit card assembly. *Proceedings of the IEEE Systems, Man, and Cybernetics Conference, Atlanta, GA, (1986)*
11. Quintana, R., Leon, V. J.: An improved group setup management strategy for pick and place SMD assembly. Working paper (1997)

Cross-Facility Production and Transportation Planning Problem with Perishable Inventory

Sandra Duni Ekşioğlu and Mingzhou Jin

Department of Industrial and Systems Engineering,
Mississippi State University, P.O. Box 9542, Mississippi State, MS 39762, USA
sde47@ie.msstate.edu

Abstract. This study addresses a production and distribution planning problem in a dynamic, two-stage supply chain. This supply chain consists of a number of facilities and retailers. The model considers that the final product is perishable and therefore has a limited shelf life. We formulate this problem as a network flow problem with a fixed charge cost function which is *NP*-hard. A primal-dual heuristic is developed that provides lower and upper bounds. The models proposed can be used for operational decisions.

1 Introduction

This paper investigates a planning model that integrates production, inventory and transportation decisions in a two-stage supply chain. Production and transportation activities have usually been studied separately by industry and academia, mainly because (i) each problem in itself is difficult and therefore the combined problem is not tractable, and (ii) different departments in an organization are in charge of each activity. In fact, the two activities can function independently if there is a sufficiently large inventory buffer that completely decouples the two. This, however, would lead to increased holding costs and longer lead times. The pressure of reducing costs in supply chains forces companies to take an integrated view of their production and distribution processes.

The supply chain analyzed in this paper consists of a number of facilities, each with similar production capabilities, and a number of retailers. We assume that retailers' demand for a single perishable product is known deterministically and that there are no production or transportation capacity constraints. Facilities produce the final product and carry inventories to satisfy retailers' demands during the planning period. We assume that there is no transshipment between facilities. This situation is typical in the food and beverage industry, where the retailers are often supermarkets and restaurants that have a very limited storage capacity. Most of the food products are perishable and have a limited lifetime. To account for this, we constraint the number of periods that a product is stored at a facility before being shipped to the retailer. The decisions that need to be made are (i) the timing of production; (ii) the location and size of inventories; and (iii) the timing of shipment.

The proposed model is suitable for tactical and operational planning. We assume that the planning period is a typical one, and it will repeat itself over

time. This means the model is cyclic in nature. For this reason, we assume a fixed starting and ending period with varying initial and ending inventories. We model the problem as a network flow problem with fixed charge cost functions. This is an *NP*-hard problem since even some of its special cases are known to be *NP*-hard. For example, the single-period problem is a fixed charge network flow problem in bipartite networks (Johnson *et al.* [1]) that is *NP*-hard. The complexity of this problem led us to consider heuristic approaches.

Production/inventory problems for perishable products have been studied. However, very little has been done in the area of integrated production and distribution planning for perishable products. Nahmias [2] presents a review of ordering policies for perishable inventories. Hsu [3] studied the economic lot-sizing problem with perishable products. In this model, the deterioration rate of the inventory and its carrying cost in each period depend on the age of the stock. Myers [4] presents a linear programming model to determine the maximum satisfiable demand for products with limited shelf life.

Previous work of the authors (Ekşioğlu *et al.* [9]) has been focused on variants of this production and distribution planning problem. In addition to previous work, this paper considers that the final product is perishable and has limited lifetime; and relaxes the assumption of constant initial inventory that is found in most inventory management models.

2 Problem Description and Formulations

This section presents a mixed integer linear programming (MILP) formulation of the cross-facility production and transportation planning problem with perishable inventory. Let F denote the number of facilities that produce and store the final products that are then delivered to R customers. The facilities have identical production capabilities. In other words, the final product can be produced in each facility. However, the setup and transportation costs, as well as the unit production and inventory holding costs, differ from one facility to the other, from one time period to the next. Given the projected demand of each retailer during the T -period planning horizon, the production and transportation planning problem decides how much to produce, transport and hold in inventory at each facility in order to meet demand at minimum cost. We assume that the planning horizon of length T is a typical one and repeats itself over time. All problem data are assumed cyclic with cycle length equal to T ($b_{j,T+1} = b_{j1}, b_{j,T+2} = b_{j2}, \dots$, where b_{jt} is the demand at retailer j in period t). As a result, the inventory pattern at the facilities will be cyclic as well. We model this by letting the initial inventory be equal to the last period inventories.

2.1 Original Problem Formulation

Property 2 of an optimal solution to our problem (Section 2.3) implies that demand in a particular time period is satisfied from exactly one facility. We develop a network flow model based on the single source assignment property. Let

p_{it} denote the unit production cost at facility i in period t ; s_{it} is the production setup cost at facility i in period t ; h_{it} is the unit inventory cost at facility i in period t ; and c_{ijt} is the total transportation cost of shipping b_{jt} from facility i to retailer j in period t . The decision variables are: q_{it} is the amount produced at facility i in period t ; I_{it} is the inventory at facility i in the end of period t ; x_{ijt} is a binary variable that equals 1 if there is a shipment from facility i to retailer j in period t , and equals 0 otherwise; and y_{it} is a binary variable that equals 1 if production takes place at facility i in period t , and equals 0 otherwise. The following is a MILP formulation of the problem:

$$\text{minimize } \sum_{i=1}^F \sum_{j=1}^R \sum_{t=1}^T \{p_{it}q_{it} + s_{it}y_{it} + h_{it}I_{it} + c_{ijt}x_{ijt}\}$$

subject to (P)

$$q_{it} + I_{i,[T+(t-1)]} - \sum_{j=1}^R b_{jt}x_{ijt} - I_{it} = 0 \quad i = 1, \dots, F; t = 1, \dots, T \quad (1)$$

$$\sum_{i=1}^F x_{ijt} = 1 \quad j = 1, \dots, R; t = 1, \dots, T \quad (2)$$

$$q_{it} - \sum_{\tau=t}^{t+k} \sum_{j=1}^R b_{j[\tau]}y_{it} \leq 0 \quad i = 1, \dots, F; t = 1, \dots, T \quad (3)$$

$$I_{it} - \sum_{\tau=t+1}^{t+k} \sum_{j=1}^R b_{j[\tau]}x_{ij[\tau]} \leq 0 \quad i = 1, \dots, F; t = 1, \dots, T \quad (4)$$

$$I_{i0} = I_{iT} \quad i = 1, \dots, F \quad (5)$$

$$q_{it}, I_{it} \geq 0 \quad i = 1, \dots, F; t = 1, \dots, T \quad (6)$$

$$y_{it}, x_{ijt} \in \{0, 1\} \quad i = 1, \dots, F; j = 1, \dots, R; t = 1, \dots, T. \quad (7)$$

For our convenience, in this formulation we have used the notation $[t] = (t+1) \bmod T - 1$ i.e., $I_{i[t-1]} = I_{i,t-1}$ for $t = 2, \dots, T$ and $I_{i[0]} = I_{iT}$.

Constraints (1) and (2) are the flow conservation constraints at the production and demand points respectively. Constraints (3) are the setup constraints. Constraints (4) are the perishability constraints, where k ($k \leq T - 1$) denotes the maximum number of periods that a product can be stored. Constraints (5) model the fact that the initial inventory is equal to the ending inventory and T is a typical sequence of periods that will repeat itself. Setting the initial inventory level equal to the ending inventory means that these inventory levels are not fixed, and the model will determine the ending inventory levels that will prepare the system for future demands. Constraints (6) are the non-negativity constraints, and (7) are the boolean constraints. Standard solvers such as CPLEX can be used to solve small instances of (P). Large problem instances are solved using the primal-dual algorithm.

The transportation cost function $f_{ij}(g_{ijt})$ is considered to be a concave function with respect to the amount shipped, g_{ijt} . Based on the single-source assignment property of an optimal solution to our problem, facility i in period t either will not ship to retailer j or will ship the total demand, b_{jt} . This indicates that the transportation cost function consists of only two points $g_{ijt} = 0$ and $g_{ijt} = b_{jt}$. The LP-relaxation of the transportation cost function passes through the points: $g_{ijt} = 0$ and $g_{ijt} = b_{jt}$. Solving the LP-relaxation of (P) with respect to the transportation cost function gives a solution such that $g_{ijt} = 0$ or $g_{ijt} = b_{jt}$. That means the LP-relaxation gives an exact approximation of the concave transportation cost function. Therefore, $c_{ijt} = f_{ij}(b_{jt})$.

In the special case when $F = 1$, retailers' demands are satisfied from the same facility; therefore, there is no decision to be made about which facility will ship the final product. In this case problem (P) reduces to the classical economic lot-sizing problem (Wagner and Whitin [5]).

2.2 Extended Problem Formulation

Linear programming relaxation of formulation (P) that is obtained by replacing the boolean constraints (7) with the nonnegativity constraints is not tight. This is due to the constraints (3). $\sum_{\tau=t}^{t+k} \sum_{j=1}^R b_{j[\tau]}$ provides a high upper bound for q_{it} , since the production in a period rarely equals this amount. One way to tighten the formulation is to split the production variables q_{it} by destination into variables $q_{ijt[\tau]}$ ($\tau = t, \dots, t + k$). The new decision variable $q_{ijt[\tau]}$ presents the amount produced at facility i in period t to satisfy the demand of retailer j in period τ . For these variables, a trivial and tight upper bound is the demand at retailer j in period τ , $b_{j\tau}$.

The following is an equivalent formulation of (P) given with respect to decision variable $q_{ijt[\tau]}$:

$$\text{minimize } \sum_{i=1}^F \sum_{t=1}^T \left[\sum_{j=1}^R \sum_{\tau=t}^{t+k} \bar{c}_{ijt[\tau]} q_{ijt[\tau]} + s_{it} y_{it} \right]$$

subject to (Ex-P)

$$\begin{aligned} \sum_{i=1}^F \sum_{t=\tau-k}^{\tau} q_{ij[T+t]\tau} &= b_{j\tau} & j &= 1, \dots, R; \tau = 1, \dots, T \\ q_{ijt[\tau]} - b_{j[\tau]} y_{it} &\leq 0 & i &= 1, \dots, F; j = 1, \dots, R; t = 1, \dots, T; t \leq \tau \leq t + k \\ q_{ijt[\tau]} &\geq 0 & i &= 1, \dots, F; j = 1, \dots, R; t = 1, \dots, T; t \leq \tau \leq t + k \\ y_{it} &\in \{0, 1\} & i &= 1, \dots, F; t = 1, \dots, T, \end{aligned}$$

where $\bar{c}_{ijt[\tau]} = p_{it} + c_{ij[\tau]}/b_{j[\tau]} + \sum_{s=t}^{\tau-1} h_{i[s]}$. In the special case when $k = 0$, no inventories are carried from one period to another. In this case the problem decomposes by period. The single-period problem is the facility location problem, which is still a difficult problem to solve.

2.3 Properties of Optimal Solution

Using the network flow interpretation, we establish the required properties of optimal solutions to (P) when the costs are nonnegative.

Theorem 1. *There exists an optimal solution to problem (P) such that the demand at retailer j in period t is satisfied from either production or the inventory of exactly one of the facilities.*

Proof: The uncapacitated, production and transportation planning problem minimizes a concave cost function over a bounded convex set; therefore, its optimal solution corresponds to a vertex of the feasible region (Zangwill [6]). Let (q^*, x^*, I^*) be an optimal solution. In an uncapacitated network flow problem, a vertex is represented by a tree solution. The tree representation of the optimal solution implies that demand in every time period will be satisfied by exactly one of the facilities (in other words, $x_{ijt}^* x_{ljt}^* = 0$, for $i \neq l$ and $t = 1, 2, \dots, T$). Furthermore, for each facility in each time period, if the inventory level is positive, there will be no production, and vice versa: $q_{it}^* I_{i,[t-1]}^* = 0$, for $i = 1, \dots, F$, $t = 1, \dots, T$. \square

Theorem 1 implies properties 1 and 2 of the optimal solutions to our problem.

Property 1. The uncapacitated, production and transportation planning problem has an optimal solution that is such that a facility in a time period t either produces or carries inventory from the previous period (or neither), but not both. This property of the optimal solutions is often referred to in the literature as the *Zero Inventory Property* (ZIP). ZIP applies to the classical single-item lot-sizing problem and some of its generalizations (Wagner and Whitin [5]).

Property 2. The optimal solution of the problem is such that the demand in a time period is satisfied from a single facility. This property is equivalent to the single-source assignment property for the uncapacitated facility location problem.

Property 3. Every facility in a given time period t either does not produce or produces the demand for a number of periods in the time interval $t, \dots, [t + k]$ (the periods do not need to be successive). This property can be easily derived from Theorem 1 and the tree representation of an optimal solution.

3 Solution Procedures

3.1 Exact Solution Approach

Theorem 2. *There exists an algorithm for the uncapacitated, single commodity, integrated production and distribution planning problem with perishable commodities (P) that is polynomial in the number of facilities and exponential in the number of periods and retailers.*

Proof: The following steps describe the algorithm:

1. Consider all the possible assignments of demands b_{jt} ($j = 1, \dots, R$ and $t = 1, \dots, T$) to facility i (for $i = 1, \dots, F$). There is a total of F^{RT} assignments.
2. Given an assignment of demands to facilities, for each facility i ($i = 1, \dots, F$) we need to solve an uncapacitated, single-commodity lot-sizing problem (ELSP _{i}) that considers the final product to be perishable and is cyclic in nature.

Without loss of optimality we can assume an optimal solution $\min_{t=1, \dots, T} I_t = 0$. We can solve problem (ELSP _{i}) for each $t = 1, \dots, T$, fixing $I_t = 0$ and treating period t as the “last” planning period. The cheapest one among the corresponding solutions is then the optimal solution. One should note that problem (ELSP _{i}) with $I_T = 0$ is in fact the (ELS) problem. This problem can be solved in $O(T \log T)$ (Wagelmans *et al.* [7]), and problem (ELSP _{i}) is solved in $O(T^2 \log T)$.

Therefore, the running time of this algorithm is bound by $O(F^{TR+1} T^2 \log T)$.

3.2 Primal-Dual Heuristic

The dual problem of the LP-relaxation of (Ex-P) has a special structure that allows us to develop a primal-dual based algorithm. The following is the formulation of the dual problem:

$$\text{maximize } \sum_{t=1}^T \sum_{j=1}^R b_{jt} v_{jt}$$

subject to (D-P)

$$\begin{aligned} \sum_{\tau=t}^{t+k} \sum_{j=1}^R b_{j[\tau]} w_{ijt[\tau]} &\leq s_{it} & i = 1, \dots, F; t = 1, \dots, T \\ v_{j[\tau]} - w_{ijt[\tau]} &\leq \bar{c}_{ijt[\tau]} & i = 1, \dots, F; t = 1, \dots, T; t \leq \tau \leq t+k \\ w_{ijt[\tau]} &\geq 0 & i = 1, \dots, F; j = 1, \dots, R; t = 1, \dots, T; t \leq \tau \leq t+k. \end{aligned}$$

In an optimal solution to (D-P), both constraints $w_{ijt[\tau]} \geq 0$ and $w_{ijt[\tau]} \geq v_{j[\tau]} - \bar{c}_{ijt[\tau]}$ should be satisfied. Since $w_{ijt[\tau]}$ is not in the objective function, we can replace it with $w_{ijt[\tau]} = \max(0, v_{j[\tau]} - \bar{c}_{ijt[\tau]})$. This leads to the following condensed dual formulation:

$$\text{maximize } \sum_{t=1}^T \sum_{j=1}^R b_{jt} v_{jt}$$

subject to (D*-P)

$$\sum_{\tau=t}^{t+k} \sum_{j=1}^R b_{j[\tau]} \max(0, v_{j[\tau]} - \bar{c}_{ijt[\tau]}) \leq s_{it} \quad i = 1, \dots, F; t = 1, \dots, T.$$

The extended formulation of the multi-facility lot-sizing problem is a special case of the uncapacitated facility location problem. The primal-dual scheme

we discuss, in principle, is similar to the primal-dual scheme proposed by Er- lenkottter [8] for the facility location problem. However, the implementation of the algorithm is different. Wagelmans *et al.* [7] use a similar primal-dual scheme for the classical lot-sizing problem. They show that this algorithm solves the problem in $O(T \log T)$. The dual variables have the following property: $v_t \geq v_{t+1}$, for $t = 1, \dots, T$. This property is used to show that the dual as- cent algorithm gives the optimal solution to the economic lot-sizing problem. This property does not hold true for (D-P).

Description of the Algorithm. Suppose the linear programming relaxation of (Ex-P) has an optimal solution (q^*, y^*) that is integral. Let (v^*, w^*) denote an optimal dual solution. The complementary slackness conditions for the primal (Ex-P) and dual (D-P) problems are as follow:

$$\begin{aligned}
 (C_1) \quad & y_{it}^* [s_{it} - \sum_{j=1}^R \sum_{\tau=t}^{t+k} b_{j[\tau]} w_{ijt[\tau]}^*] = 0 \text{ for } i = 1, \dots, F; t = 1, \dots, T \\
 (C_2) \quad & q_{ijt[\tau]}^* [\bar{c}_{ijt[\tau]} - v_{j[\tau]}^* + w_{ijt[\tau]}^*] = 0 \text{ for } i = 1, \dots, F; j = 1, \dots, R; \\
 & \quad \quad \quad t = 1, \dots, T; t \leq \tau \leq t + k \\
 (C_3) \quad & w_{ijt[\tau]}^* [q_{ijt[\tau]}^* - b_{j[\tau]} y_{it}^*] = 0 \text{ for } i = 1, \dots, F; j = 1, \dots, R; \\
 & \quad \quad \quad t = 1, \dots, T; t \leq \tau \leq t + k \\
 (C_4) \quad & v_{jt}^* [b_{jt} - \sum_i \sum_{\tau=t-k}^t q_{ijt[T+\tau]}^*] = 0 \text{ for } j = 1, \dots, R; t = 1, \dots, T.
 \end{aligned}$$

The simple structure of the dual problem can be exploited to obtain near optimal feasible solutions by inspection. Suppose that the optimal values of the first $f - 1$ dual variables of (D*-P) are known. Then, to be feasible, the f -th dual variable ($v_{l\tau}$) must satisfy the following constraints:

$$\begin{aligned}
 & b_{l\tau} \max(0, v_{l\tau} - \bar{c}_{ill\tau}) \leq M_{ill,\tau-1} = s_{it} - \\
 & \sum_{j=1}^R \sum_{s=t}^{\tau-1} b_{j[T+s]} \max(0, v_{j[T+s]} - \bar{c}_{ij[T+t][T+s]}) - \sum_{j=1}^l b_{j\tau} \max(0, v_{j\tau} - \bar{c}_{ij[T+t]\tau})
 \end{aligned}$$

for all $i = 1, \dots, F$ and $t = \tau - k, \dots, \tau$. In order to maximize the dual problem, we should assign $v_{l\tau}$ the largest value satisfying these constraints. When $b_{l\tau} > 0$, this value is

$$v_{l\tau} = \min_{i=1, \dots, F; \tau \geq t} \left\{ \bar{c}_{ill\tau} + \frac{M_{ill,\tau-1}}{b_{l\tau}} \right\} \tag{8}$$

Note that if $M_{ill\tau-1} \geq 0$ implies $v_{l\tau} \geq \bar{c}_{ill\tau}$.

A dual feasible solution can be obtained simply by calculating the value of the dual variables sequentially (Figure 1). A backward construction algorithm can then be used to generate primal feasible solutions (Figure 2). The primal-dual solutions found using these algorithms may not necessarily satisfy the comple- mentary slackness conditions.

Theorem 3. *The solutions obtained with the primal and dual algorithms are feasible and they always satisfy the complementary slackness conditions (C₁) and (C₂).*

Proof: The proof is similar to the proof of Proposition 4.1 in Ekşioğlu *et al.* [9].

Hence, one can determine whether the solution obtained with the primal and dual algorithms is optimal by checking if conditions (C₃) are satisfied or if the

```

 $M_{i\tau, \tau-1} = s_{i\tau}$  for  $i = 1, \dots, F; j = 1, \dots, R; \tau = 1, \dots, T$ 
for  $\tau = 1$  to  $T$  do
  for  $j = 1$  to  $R$  do
    if  $b_{j\tau} = 0$  then  $v_{j\tau} = 0$ 
    else
       $v_{j\tau} = \min_{it} \{ \bar{c}_{ij[T+t]\tau} + M_{i[T+t], \tau-1} / b_{j\tau} \}; \tau - k \leq t \leq \tau$ 
    for  $t = \tau - k$  to  $\tau$  do
      for  $i = 1$  to  $F$  do
         $M_{i[T+t]\tau} = \max\{0, M_{i[T+t], \tau-1} - b_{j\tau} * \max\{0, v_{j\tau} - \bar{c}_{ij[T+t]\tau}\}\}$ 
      enddo
    enddo
  enddo
enddo

```

Fig. 1. Dual algorithm

```

 $y_{it} = 0, q_{ijt\tau} = 0, i = 1, \dots, F; j = 1, \dots, R; t = 1, \dots, T; \tau \leq [t + k]$ 
 $P = \{(j, l) | b_{jl} > 0, \text{ for } j = 1, \dots, R; l = 1, \dots, T\}$ 
Start :  $\tau = \max l \in P, t = \tau - k$ 
Step 1 : for  $i = 1$  to  $F$  do
  for  $j = 1$  to  $R$  do
    repeat  $t = t + 1$ 
      until  $M_{i[T+t]\tau} = 0$  and  $\bar{c}_{ij[T+t]\tau} - v_{j\tau} + \max\{0, v_{j\tau} - \bar{c}_{ij[T+t]\tau}\} = 0$ 
       $y_{i[T+t]\tau} = 1$ , and  $i = i, t = t$ , go to Step 2
    enddo
  enddo
  go to Step 3
Step 2 : for  $t = t$  to  $t + k$  do
  for  $j = 0$  to  $R$  do
    if  $\bar{c}_{i j t [t]} - v_{j[t]} + \max\{0, v_{j[t]} - \bar{c}_{i j t [t]}\} = 0$ 
      then  $q_{i j t [t]} = b_{j[t]}, P = P - (j, [t])$ 
    enddo
  enddo
Step 3 : if  $P \neq \emptyset$  then go to Start

```

Fig. 2. Primal algorithm

objective function values from the primal and dual algorithms are equal. The running time of this primal-dual algorithm is $O(FRT^2)$.

4 Computational Results

In this section we describe our computational experience in solving the integrated production and transportation planning problem with perishable inventory. Our

Table 1. Summary of results of primal-dual algorithm

Problem	Setup Costs									
	200-300		200-900		600-900		900-1,500		1,200-1,500	
	Error (%)	Cpu (sec)	Error (%)	Cpu (sec)	Error (%)	Cpu (sec)	Error (%)	Cpu (sec)	Error (%)	Cpu (sec)
16	0.14	0.15	0.24	0.15	0.77	0.15	1.37	0.15	1.91	0.15
17	0.19	0.15	0.30	0.20	1.08	0.20	1.91	0.20	2.64	0.20
18	0.26	0.20	0.37	0.25	1.30	0.20	2.28	0.25	3.21	0.25
19	0.14	0.70	0.23	0.55	0.77	0.65	1.40	0.65	1.96	0.06
20	0.19	1.00	0.29	0.85	1.07	0.95	1.85	0.95	2.59	0.95
21	0.24	1.25	0.34	1.25	1.30	1.25	2.26	1.25	3.17	1.25
22	0.14	2.95	0.23	2.90	0.78	2.85	1.40	3.10	1.94	2.95
23	0.19	4.45	0.30	4.45	1.08	4.45	1.89	5.00	2.63	4.95
24	0.24	6.00	0.35	5.85	1.29	6.05	2.23	6.10	3.12	5.80

goal is to provide some indication of both the quality and the computing time of the lower and upper bounds generated using the primal-dual algorithm. The problem instances we use are the same as the ones presented in Ekşioğlu *et al.* [9]. For all problems we set $k = \frac{T}{2}$. The errors presented are calculated as follows:

$$Error(\%) = \frac{\text{Primal Sol.} - \text{Dual Sol.}}{\text{Dual Sol.}} * 100.$$

It has been shown in the literature (Hochbaum and Segev [10], Ekşioğlu *et al.* [9]) that the ratio of setup to variable costs impacts the complexity of the fixed charge network flow problems. The results in Table 1 indicate that an increase in setup costs impacts the quality of the solutions from primal-dual algorithm. However, setup costs do not affect the running time of the algorithm. The results also show that an increase in the number of facilities and time periods impacts the performance of the primal-dual algorithm.

For the same set of problems, formulation (Ex-P) is solved using CPLEX 9 callable libraries. CPLEX gives the optimal solution for problems 16 to 18, but fails to solve the rest of the problems because of the problem size. Table 2 presents the running time of CPLEX for problems 16 to 18.

Table 2. CPLEX running times (in cpu seconds)

Problem	Setup Costs				
	200-300	200-900	600-900	900-1,500	1,200-1,500
16	15.85	15.95	16.05	16.55	16.50
17	26.55	27.05	28.35	28.75	44.40
18	40.25	41.05	42.20	42.10	86.50

5 Conclusions

This paper presents two network flow formulations for an integrated production and transportation planning problem with perishable inventories. The network consists of a number of facilities and retailers. The facilities produce and carry inventory to satisfy retailers' demands during T time periods. Retailers' demands are known deterministically. Unlike the traditional inventory models, the starting and ending inventories are not constant. Section 3 presents an exact solution procedure and a primal-dual algorithm to solve the problem. The exact algorithm is polynomial in the number of facilities and exponential in the number of retailers and time periods. This algorithm runs in $O(F^{TR+1}T^2 \log T)$ times. The primal-dual algorithm is used to generate lower and upper bounds. Its running time is $O(FRT^2)$ times. We tested the performance of the algorithms on a wide range of randomly generated problems. Computational results show high-quality solutions from the primal-dual algorithm. The maximum error gap was 3.12 percent and the maximum running time was 6.10 cpu seconds. Computational results demonstrate that the ratio of fixed to variable costs, the length of time horizon and the number of facilities impacted the running time and the quality of the solutions from the primal-dual algorithm and CPLEX. We identified a number of problems that CPLEX could not solve because it ran out of memory. However, for all problem classes, primal-dual algorithm gave high-quality solutions in a reasonable amount of time.

References

1. Johnson, D.S., Lenstra, J.K., Rinnooy Kan, A.H.G.: The complexity of the network design problem. *Networks* **8** (1978) 279–285
2. Nahmias, S.: Perishable inventory theory: A review. *Operations Research* **30** (1982) 680–708
3. Hsu, V.N.: Dynamic economic lot size model with perishable inventory. *Management Science* **46** (2000) 1159–1169
4. Myers, D.C.: Meeting seasonal demand for products with limited shelf lives. *Naval Research Logistics* **44** (1997) 473–483
5. Wagner, H.M., Whitin, T.M.: Dynamic version of the economic lot size model. *Management Science* **5** (1958) 89–96
6. Zangwill, W.I.: Minimum concave cost flows in certain networks. *Management Science* **14** (1968) 429–450
7. Wagelmans, A., van Hoesel, S., Kolen, A.: Economic lot sizing: An $O(n \log n)$ algorithm that runs in linear time in the wagner-whitin case. *Operations Research* **40** (1992) 145–156
8. Erlenkotter, D.: A dual-based procedure for uncapacitated facility location. *Operations Research* **26** (1978) 992–1009
9. Ekşioğlu, S.D., Romeijn, H.E., Pardalos, P.M.: Cross-facility management of production and transportation planning problem. *Comp. Oper. Res.* (in press)
10. Hochbaum, D.S., Segev, A.: Analysis of a flow problem with fixed charges. *Networks* **19** (1989) 291–312

A Unified Framework for the Analysis of $M/G/1$ Queue Controlled by Workload

Ho Woo Lee¹, Se Won Lee¹, Won Ju Seo¹,
Sahng Hoon Cheon¹, and Jongwoo Jeon²

¹ Department of Systems Management Engineering,
Sungkyunkwan University, Su Won, Korea 440-746
hwlee@skku.edu

<http://web.skku.edu/~or>

² Department of Statistics,
Seoul National University, Seoul, Korea 151-742
jwjeon@plaza.snu.ac.kr

Abstract. In this paper, we develop a unified framework for the analysis of the waiting time, the sojourn time and the queue length of the $M/G/1$ queue in which the server is controlled by workload. We apply our framework to the D -policy $M/G/1$ queue and confirm the results that already exist in the literature. We also use our approach and derive, for the first time, the sojourn time distribution of an arbitrary customer in the $M/G/1$ queue under D -policy. The methodologies developed in this paper can be applied to a wide range of $M/G/1$ queueing systems in which the server is controlled by workload.

1 Introduction

In the D -policy queueing system, the server is turned off as soon as the system becomes empty. It is reactivated only when the cumulative workload (i.e. sum of the service times) exceeds some predetermined threshold D . It is noted that the service times of the customers who arrive during the idle period are neither identical nor independent.

The pioneering studies on the D -policy queue were carried out by Balachandran [2], Balachandran and Tijms [3], Boxma [4] and Tijms [22]. Their primary concern was in the optimal control of D under a linear cost structure by using the mean workload.

Studies on the queue length of the D -policy queue could not be found until Dshalalow [9] studied the queue length process of the D -policy queue with vacations. But as pointed out by Artalejo [1] and Chae and Park [7][8], his D -policy did not agree with the classical D -policy in the true sense because he implicitly assumed that the customers who have arrived during the idle period are assigned totally new iid service times when the busy period begins.

The first successful works on the queue length of the D -policy queue in the true sense were carried out by Chae and Park [8] and Artalejo [1]. Their results show that the well-known decomposition property of the $M/G/1$ queue with

generalized vacations does not hold for the D -policy queueing system due to the dependencies of the service times. Studies on the D -policy queueing system with MAP (Markovian Arrival process) arrivals were explored by Lee and Song [14] and Lee et al. [13]. More works concerning D -policy can be found in Sivazlian [20], Li and Niu [15], Rhee [18], Park and Chae [17], Lillo and Martin [16], Feinberg and Kella [10], Lee and Baek [11] and Lee et al. [12].

The first studies on the waiting time and queue length of the $M/G/1/D$ -policy queue took different approaches. Park and Chae [17] used the supplementary variable technique to obtain the waiting time distribution. But Artalejo [1] used the concept of the regeneration cycle to derive the exact queue length distribution. Chae and Park [8] derived the PGF (probability generating function) of the queue length distribution by analyzing the departure points.

Our objective in this paper is twofold:

- (a) We develop a methodology to analyze the queue length, waiting time and sojourn time under a unified framework.
- (b) We use our framework to confirm the existing results. We also derive the sojourn time distribution under the unified methodology. This sojourn time distribution will be the first one derived in the queueing literature.

We note that the methodologies developed in this paper can be applied to a wide range of queueing systems in which the server is controlled by workload.

2 Preliminary Work and Case Classification

In this section, we lay the foundations for the unified framework that will be presented in subsequent sections. The starting point is to analyze the workload process during the idle period. The rationale is to get the information of the workload and the remaining idle period at an arbitrary arrival time during the idle period. These two quantities affect the queue length and waiting times of the future process.

Before going into the details, let us define some notation. As for the service time random variable S , we will use $s(x)$, $S(x)$, $s^{(n)}(x)$ and $S^{(n)}(x)$ as the pdf, distribution function (DF), n -fold pdf and n -fold DF respectively. If we define \mathfrak{R} as the renewal process that is generated by iid service times, then

$$M(x) = \sum_{n=1}^{\infty} S^{(n)}(x) \quad \text{and} \quad m(x) = \frac{d}{dx}M(x) = \sum_{n=1}^{\infty} s^{(n)}(x) \tag{2.1}$$

are the renewal function and the renewal density function of the \mathfrak{R} process.

Let us define $I(x, y)$ as the indicator random variable that takes 1 if the workload process during the idle period passes through the interval $(x, y]$ and 0 otherwise. Let us define its probability

$$\phi(x)dx = Pr[I(x, x + dx) = 1], \quad (x > 0). \tag{2.2}$$

The event $\{I(x, x + dx) = 1\}$ occurs iff one event of the \mathfrak{R} -process occurs during $(x, x + dx]$. Thus, we have

$$\phi(x) = m(x). \tag{2.3}$$

Let $U_{idle}(x)$ be the probability that the workload at an arbitrary time during the idle period is less than or equal to x with $u_{idle}(x) = \frac{d}{dx}U_{idle}(x)$, $(0 < x \leq D)$ as its pdf. Then, we have

$$U_{idle}(0) = \frac{1}{E(N_D)}, \tag{2.4a}$$

$$u_{idle}(x) = \frac{m(x)}{E(N_D)}, \quad (0 < x \leq D), \tag{2.4b}$$

where N_D is the number of customers at the busy period starting point and I_D is the length of the idle period.

For N_D , we have (Ross [19]),

$$Pr(N_D = k) = S^{(k-1)}(D) - S^{(k)}(D). \tag{2.5}$$

Then, we easily get the PGF and the mean as

$$\begin{aligned} N_D(z) &= \sum_{k=1}^{\infty} z^k Pr(N_D = k) = \sum_{k=1}^{\infty} z^k [S^{(k-1)}(D) - S^{(k)}(D)] \\ &= 1 - (1 - z) \sum_{k=0}^{\infty} z^k S^{(k)}(D), \end{aligned} \tag{2.6a}$$

$$E(N_D) = \frac{d}{dz}N_D(z)\Big|_{z=1} = 1 + M(D). \tag{2.6b}$$

Let U_D be the workload at the busy period starting point. If we condition on the amount of work just before D , the pdf of U_D becomes

$$u_D(x) = s(x) + \int_0^D s(x - y)m(y)dy, \quad (x > D) \tag{2.7a}$$

with the Laplace-Stieltjes transform (LST)

$$\begin{aligned} U_D^*(\theta) &= E(e^{-\theta U_D}) = \int_D^{\infty} e^{-\theta x}u_D(x)dx \\ &= 1 - [1 - S^*(\theta)] \left[1 + \int_0^D e^{-\theta x}m(x)dx \right]. \end{aligned} \tag{2.7b}$$

Then we have

$$E(U_D) = -\frac{d}{d\theta}U_D^*(\theta)\Big|_{\theta=0} = E(N_D)E(S) = E(S)[1 + M(D)]. \tag{2.7c}$$

Let us categorize the customers into two types as follows:

- SC(Special Customer) : the customer who arrives during the idle period.
- OC(Ordinary Customer) : the customer who arrives when the server is busy.

The workload and the remaining idle period at the arrival instance of an arbitrary SC (test-SC) determines the queue length and waiting time of the customers who arrive thereafter. Thus it is important to classify the possible situations at its arrival instance. We have four cases (Figure 1).

- (1) The test-SC is the first customer during the idle period.
 - (Case 1) The total work just after its arrival is less than or equal to D (Ⓐ) in Figure 1).
 - (Case 2) The total work just after its arrival is greater than D (Ⓓ).
- (2) The test-SC is not the first customer during the idle period.
 - (Case 3) The total work just after its arrival is less than or equal to D (Ⓑ).
 - (Case 4) The total work just after its arrival is greater than D (Ⓒ).

Above case classification is important since the derivations of the distributions of the queue length, the waiting time and the sojourn time in subsequent sections will be based on the above four cases.

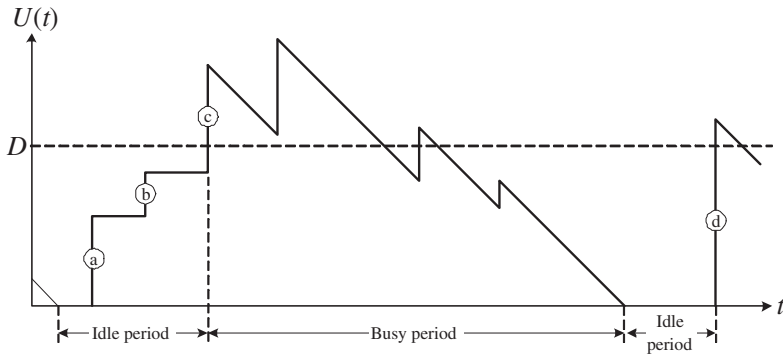


Fig. 1. Four cases at the arrival instance of the test-SC

3 Queue Length

In this section, we derive the queue length distribution based on the case classification. We first note that the queue length PGFs $\bar{\Pi}(z)$, $\Pi(z)$ and $P(z)$ at an arrival, at a departure and at an arbitrary time point are all equal due to PASTA (Wolff [23]) and we have

$$\bar{\Pi}(z) = \Pi(z) = P(z). \tag{3.1}$$

Thus, we just need to find the departure point PGF $\Pi(z)$.

Since the work is still conserved in our system, an arbitrary customer is an OC with probability ρ and a SC with probability $(1 - \rho)$. Thus, conditioning on the customer type, we have

$$P(z) = \Pi(z) = \rho I_{oc}(z) + (1 - \rho) I_{sc}(z), \tag{3.2}$$

where $I_{oc}(z)$ is the PGF of the queue length at an arbitrary OC departure and $I_{sc}(z)$ is the PGF of the queue length at an arbitrary SC departure.

To derive the queue length PGF $\Pi_{oc}(z)$, we note that the departure process of the OCs is stochastically equivalent to the departure process of the $M/G/1$ queue that starts with N_D^{**} customers where N_D^{**} is the number of customers that arrive during U_D (the workload at the start of the busy period). Thus from the well-known decomposition property of the $M/G/1$ queue with generalized vacations (Takagi [21]), we get

$$\Pi_{oc}(z) = \Pi_{M/G/1}(z) \cdot N_E^{**}(z) = \frac{(1 - \rho)(1 - z)S^*(\lambda - \lambda z)}{S^*(\lambda - \lambda z) - z} \cdot \frac{1 - U_D^*(\lambda - \lambda z)}{\lambda E(U_D)(1 - z)}, \tag{3.3}$$

where $\Pi_{M/G/1}(z)$ is the queue length PGF in the ordinary $M/G/1$ queue and $N_E^{**}(z)$ is the PGF of the backward recurrence time of the discrete-time renewal process that is generated by iid N_D^{**} 's.

The derivation of $\Pi_{sc}(z)$ depends on the four cases we mentioned in Section 2. First, the queue length left behind by the departing test-SC is the sum of the following two:

- (i) the number of OCs who arrive during the workload S_T just after its arrival (including its service time),
- (ii) the number N_R of the OCs who arrive behind the test-SC during the idle period.

Let us define the joint distribution of S_T and N_R as

$$\alpha(x, n)dx = Pr(x < S_T \leq x + dx, N_R = n), \quad (x \geq 0, n \geq 0). \tag{3.4}$$

We note that the queueing process after the arrival of the test-SC is stochastically equivalent to the idle period under $(D - x)$ -policy.

Now, we have

$$\Pi_{sc}(z) = \int_0^\infty e^{-(\lambda - \lambda z)x} \sum_{n=0}^\infty z^n \alpha(x, n)dx. \tag{3.5}$$

Let $\alpha_{(i)}(x, n)$ be the value of $\alpha(x, n)$ under case i mentioned in Section 2. Then, we get

(Case 1) $\alpha_{(1)}(x, n) = U_{idle}(0)s(x)\psi_n^{D-x}, \quad (n \geq 1, 0 < x \leq D),$ (3.6a)

(Case 2) $\alpha_{(2)}(x, n) = U_{idle}(0)s(x), \quad (n = 0, x > D),$ (3.6b)

(Case 3) $\alpha_{(3)}(x, n) = \left[\int_{y=0+}^x u_{idle}(y)s(x-y)dy \right] \psi_n^{D-x}, \quad (n \geq 1, 0 < x \leq D),$ (3.6c)

(Case 4) $\alpha_{(4)}(x, n) = \int_{y=0+}^D u_{idle}(y)s(x-y)dy, \quad (n = 0, x > D),$ (3.6d)

where $\psi_n^v = Pr(N_v = n)$ is the probability that the busy period starts with n customers under v -policy which is given by, from (2.5),

$$\psi_n^v = S^{(n-1)}(v) - S^{(n)}(v), \quad (n \geq 1). \tag{3.7}$$

If we use (3.6a-d) in (3.5) after using (2.4a,b), we get

$$\Pi_{sc}(z) = \frac{1}{E(N_D)} \left[U_D^*(\lambda - \lambda z) + \int_0^D e^{-(\lambda - \lambda z)x} \sum_{n=1}^{\infty} z^n \psi_n^{D-x} m(x) dx \right], \tag{3.8}$$

where $U_D^*(\theta)$ is the LST of the DF of U_D .

Now, using (3.3) and (3.8) in (3.2) yields

$$\begin{aligned} P(z) &= \frac{1 - \rho}{E(N_D)} \\ &\times \left\{ U_D^*(\lambda - \lambda z) + \sum_{n=1}^{\infty} z^n \int_0^D e^{-(\lambda - \lambda z)x} [S^{(n-1)}(D - x) - S^{(n)}(D - x)] m(x) dx \right\} \\ &+ \rho \cdot \frac{(1 - \rho)(1 - z)S^*(\lambda - \lambda z)}{S^*(\lambda - \lambda z) - z} \cdot \frac{[1 - U_D^*(\lambda - \lambda z)]}{\lambda E(U_D)(1 - z)}. \end{aligned} \tag{3.9}$$

We confirm that above $P(z)$ is exactly equal to the one that was derived by Chae and Park [8] by using different approach.

4 Waiting Time

Conditioning on the customer type again, we have

$$W_q^*(\theta) = (1 - \rho)W_{q,sc}^*(\theta) + \rho W_{q,oc}^*(\theta). \tag{4.1}$$

Noting that the busy period is a delay cycle with initial delay U_D , we have (Takagi [21])

$$W_{q,oc}^*(\theta) = W_{q,M/G/1}^*(\theta) \cdot U_{D,E}^*(\theta) = \frac{(1 - \rho)\theta}{\theta - \lambda + \lambda S^*(\theta)} \cdot \frac{1 - U_D^*(\theta)}{\theta E(U_D)}, \tag{4.2}$$

where $U_{D,E}^*(\theta)$ is the LST of the DF of the forward recurrence time of U_D .

$W_{q,sc}^*(\theta)$ can be obtained based on the four cases mentioned in Section 2. Let $W_{q,sc(i)}^*(\theta)$ be the LST of the waiting time in case i . Then, we have

$$\text{(Case 1)} \quad W_{q,sc(1)}^*(\theta) = \int_{x=0}^D U_{idle}(0) s(x) I_{D-x}^*(\theta) dx, \tag{4.3a}$$

$$\text{(Case 2)} \quad W_{q,sc(2)}^*(\theta) = U_{idle}(0) [1 - S(D)], \tag{4.3b}$$

$$\text{(Case 3)} \quad W_{q,sc(3)}^*(\theta) = \int_{x=0}^D \int_{y=0+}^x e^{-\theta y} u_{idle}(y) s(x - y) I_{D-x}^*(\theta) dy dx, \tag{4.3c}$$

$$\text{(Case 4)} \quad W_{q,sc(4)}^*(\theta) = \int_{x=D}^{\infty} \int_{y=0+}^D e^{-\theta y} u_{idle}(y) s(x - y) dy dx, \tag{4.3d}$$

where $I_{D-x}^*(\theta)$ is the length of the idle period with threshold $D - x$ and

$$I_{D-x}^*(\theta) = \sum_{n=1}^{\infty} \left(\frac{\lambda}{\theta + \lambda} \right)^n \left[S^{(n-1)}(D - x) - S^{(n)}(D - x) \right]. \tag{4.4}$$

Using (2.4a,b) and (4.4) in (4.3a-d) after using (2.4a,b), we get

$$W_{q,sc}^*(\theta) = \sum_{i=1}^4 W_{q,sc(i)}^*(\theta) = \frac{1}{E(N_D)} \left\{ \sum_{k=0}^{\infty} \left(\frac{\lambda}{\theta + \lambda} \right)^k \left[S^{(k)}(D) - S^{(k+1)}(D) \right] + \int_0^D e^{-\theta x} \sum_{k=0}^{\infty} \left(\frac{\lambda}{\theta + \lambda} \right)^k \left[S^{(k)}(D - x) - S^{(k+1)}(D - x) \right] m(x) dx \right\}. \tag{4.5}$$

Now, using (4.2) and (4.5) in (4.1) yields

$$W_q^*(\theta) = U^*(\theta) + \frac{1 - \rho}{E(N_D)} \left[\left(\frac{\lambda}{\theta + \lambda} \right) - 1 \right] \times \left[\sum_{k=1}^{\infty} \left(\frac{\lambda}{\theta + \lambda} \right)^{k-1} S^{(k)}(D) + \int_0^D e^{-\theta x} \sum_{k=1}^{\infty} \left(\frac{\lambda}{\theta + \lambda} \right)^{k-1} S^{(k)}(D - x) m(x) dx \right]. \tag{4.6}$$

It is confirmed that this is exactly equal to the one which was derived by using the supplementary variable technique by Park and Chae [17].

5 System Sojourn Time

In the usual $M/G/1$ queueing system without D -policy, the sojourn time is the simple sum of the two independent random variables, the waiting time and the service time:

$$W_{M/G/1}^*(\theta) = W_{q,M/G/1}^*(\theta) \cdot S^*(\theta). \tag{5.1}$$

But (5.1) is no longer true for the D -policy queueing system because the waiting time and the service time are dependent.

Conditioning on the customer type, the LST of the sojourn time DF can be written as

$$W^*(\theta) = (1 - \rho)W_{sc}^*(\theta) + \rho W_{oc}^*(\theta). \tag{5.2}$$

Since the waiting time and the service of an arbitrary OC are independent, we have from (4.2),

$$W_{oc}^*(\theta) = W_{M/G/1}^*(\theta) \cdot U_{D,E}^*(\theta) = \frac{(1 - \rho)\theta S^*(\theta)}{\theta - \lambda + \lambda S^*(\theta)} \cdot \frac{1 - U_D^*(\theta)}{\theta E(U_D)}. \tag{5.3}$$

Let $W_{sc(i)}^*(\theta)$ be the LST of the sojourn time in case i . Then, we have

$$\text{(Case 1)} \quad W_{sc(1)}^*(\theta) = \int_{x=0}^D e^{-\theta x} U_{idle}(0) s(x) I_{D-x}^*(\theta) dx, \tag{5.4a}$$

$$\text{(Case 2)} \quad W_{sc(2)}^*(\theta) = \int_{x=D}^{\infty} e^{-\theta x} U_{idle}(0) s(x) dx, \tag{5.4b}$$

$$\text{(Case 3)} \quad W_{sc(3)}^*(\theta) = \int_{x=0}^D e^{-\theta x} \int_{y=0+}^x u_{idle}(y) s(x-y) I_{D-x}^*(\theta) dy dx, \tag{5.4c}$$

$$\text{(Case 4)} \quad W_{sc(4)}^*(\theta) = \int_{x=D}^{\infty} e^{-\theta x} \int_{y=0+}^D u_{idle}(y) s(x-y) dy dx, \tag{5.4d}$$

where $I_{D-x}^*(\theta)$ was given in (4.4).

Combining all four cases, we get, after using (2.4a,b),

$$\begin{aligned} W_{sc}^*(\theta) &= \sum_{i=1}^4 W_{sc(i)}^*(\theta) = \frac{1}{E(N_D)} \\ &\times \left\{ U_D^*(\theta) + \int_0^D e^{-\theta x} \sum_{n=1}^{\infty} \left(\frac{\lambda}{\theta + \lambda} \right)^n \left[S^{(n-1)}(D-x) - S^{(n)}(D-x) \right] m(x) dx \right\}. \end{aligned} \tag{5.5}$$

Using (5.3) and (5.5) in (5.2), we get

$$\begin{aligned} W^*(\theta) &= \frac{1}{E(N_D)} \\ &\times \left\{ U_D^*(\theta) + \int_0^D e^{-\theta x} \sum_{n=1}^{\infty} \left(\frac{\lambda}{\theta + \lambda} \right)^n \left[S^{(n-1)}(D-x) - S^{(n)}(D-x) \right] m(x) dx \right. \\ &\quad \left. + \frac{\lambda S^*(\theta) [1 - U_D^*(\theta)]}{\theta - \lambda + \lambda S^*(\theta)} \right\}. \end{aligned} \tag{5.6}$$

We note that above LST of the system sojourn time is derived for the first time in this paper.

The mean sojourn time becomes

$$\begin{aligned} W &= -\frac{d}{d\theta} W^*(\theta) \Big|_{\theta=0} \\ &= E(S) + \frac{\lambda E(S^2)}{2(1-\rho)} + \frac{\int_0^D x \cdot m(x) dx}{E(N_D)} + \frac{1-\rho}{\lambda E(N_D)} \int_0^D [1 + M(D-x)] m(x) dx \\ &= W_q + E(S). \end{aligned} \tag{5.7}$$

6 Summary

In this paper, we presented a unified methodology that can be used to analyze the queue length, the waiting time and the system sojourn time under the same

framework. The framework is based on the case classification of the workload at the arrival instance of an arbitrary special customer. We confirmed results that exist in the literature which had been derived by using different approaches. The sojourn time distribution we derived by using our approach was the first one in the queueing literature.

Our approach can be extensively applied to analyze a variety of $M/G/1$ queueing systems in which the server is controlled by workload.

Acknowledgement. This work was supported by the SRC/ERC program of MOST/KOSEF (grant R11-2000-073-00000).

References

1. Artalejo, J.R. : On the $M/G/1$ queue with D -policy. Applied Mathematical Modelling **25** (2001) 1055-1069
2. Balachandran, K.R. : Control policies for a single server system, Management Sci. **19** (1973) 1013-1018
3. Balachandran, K.R. and Tijms, H. : On the D -policy for the $M/G/1$ queue, Management Sci. **21(9)** (1975) 1073-1076
4. Boxma, O.J. : Note on a control problem of Balachandran and Tijms, Management Sci. **22(8)** (1976) 916-917
5. Boxma, O.J. : Workloads and waiting times in single-server systems with multiple customer classes, Queueing Systems **5** (1989) Nos.1-3 185-214
6. Boxma, O.J. and Groenendijk, W.P. : Pseudo-conservation laws in cyclic-service systems, J. Appl. Probab. **24(4)** (1987) 949-964
7. Chae, K.C. and Park, Y.I. : On the optimal D -policy for the $M/G/1$ queue, J Korean Institute of Industrial Engineers (KIIE) **25(4)** (1999) 527-531
8. Chae, K.C. and Park, Y.I. : The queue length distribution for the $M/G/1$ queue under the D -policy, J. Appl. Prob. **38(1)** (2001) 278-279
9. Dshalalow, J.H. : Queueing processes in bulk systems under the D -policy, J. Appl. Probab. **35** (1998) 976-989
10. Feinberg, U.A. and Kella, O. : Optimality of D -policies for an $M/G/1$ queue with a removable server, Queueing Systems **42** (2002) 355-376
11. Lee, H.W. and Baek, J.W. : BMAP/ $G/1$ queue under D -policy: queue length analysis, Stochastic Models **21(2-3)** (2005) 1-21
12. Lee, H.W., Baek, J.W. and Jeon, J. : Analysis of queue under D -policy, Stochastic Analysis and Applications **23** (2005) 785-808
13. Lee, H.W., Cheon, S.H., Lee, E.Y. and Chae, K.C. : Workload and waiting time analysis of MAP/ $G/1$ queue under D -policy, Queueing Systems **48** (2004) 421-443
14. Lee, H.W. and Song, K.S. : Queue length analysis of MAP/ $G/1$ queue under D -policy, Stochastic Models **20(3)** (2004) 363-380
15. Li, J. and Niu, S.C. : The waiting time distribution for the $GI/G/1$ queue under the D -policy, Prob. Eng. Inf. Sci. **6** (1992) 287-308
16. Lillo, R.E. and Martin, M. : On optimal exhaustive policies for the $M/G/1$ queue, Operations Research Letters **27** (2000) 39-46
17. Park, Y.I. and Chae, K.C. : Analysis of unfinished work and queue waiting time for the $M/G/1$ queue with D -policy, Journal of the Korean Statistical Society **28(4)** (1999) 523-533

18. Rhee, H.K. : Development of a new methodology to find the expected busy periods for controllable M/G/1 queueing models operating under the multi-variable operating policies : concepts and applications to the dyadic policies, J Korean Institute of Industrial Engineers (KIIE) **23(4)** (1997) 729-739
19. Ross, S.M : Stochastic Processes, 2nd ed., John Wiley & Sons, Inc. (1996)
20. Sivazlian, B.D. : Approximate optimal solution for a D -policy in an M/G/1 queueing system, AIIE Transactions **11** (1979) 341-343.
21. Takagi, H. : Queueing Analysis: A Foundation of Performance Evaluation, Vol. I, Vacation and Priority Systems, Part I. North-Holland: Amsterdam (1991)
22. Tijms, H.C., : Optimal control of the workload in an M/G/1 queueing system with removable server, Math. Operationsforsch. u. Statist. **7** (1976) 933-943
23. Wolff, R.W., : Poisson arrivals see time averages, Oper. Res. **30(2)** (1982) 223-231

Tabu Search Heuristics for Parallel Machine Scheduling with Sequence-Dependent Setup and Ready Times

Sang-Il Kim, Hyun-Seon Choi, and Dong-Ho Lee

Department of Industrial Engineering, Hanyang University,
Sungdong-gu, Seoul 133-791, Korea
leman@hanyang.ac.kr

Abstract. We consider the problem of scheduling a set of independent jobs on parallel machines for the objective of minimizing total tardiness. Each job may have sequence-dependent setup and distinct ready times, i.e., the time at which the job is available for processing. Due to the complexity of the problem, tabu search heuristics are suggested that incorporate new methods to generate the neighborhood solutions. Three initial solution methods are also suggested. Computational experiments are done on a number of randomly generated test problems, and the results show that the tabu search heuristics suggested in this paper outperform the existing one.

1 Introduction

Occurrence of machines in parallel is common in various manufacturing and assembly systems. For example, various parallel machine systems can be found in semiconductor manufacturing, printed circuit board (PCB) manufacturing, group technology (GT) cell with identical machines, steel manufacturing, injection molding process, etc. Among them, the injection molding process, which has a huge global market, is forced to reduce costs as well as increase productivity, and efficient scheduling is an important means to achieve them (Dastidar and Nagi 2004).

In general, the parallel machine scheduling problem has two distinct decisions: allocation and sequencing. Allocation is a decision concerning the assignment of jobs to machines, while sequencing is to order the jobs assigned to each machine. Also, parallel machine scheduling can be classified according to: a) objective; b) machine type (identical or non-identical machine); c) job type (independent or dependent job); and d) preemption. Additional criteria for the problem classification are sequence-dependent setup times and ready times. Here, the sequence-dependent setups imply that setup times depend on the type of job just completed and on the job to be processed. An example of a process with sequence dependent setup times can be found in the injection molding process. In this example, only the mold change time is required if the same material is used between two consecutive jobs, while the screw cleaning times are additionally required, otherwise. Also, the ready time of a job, which is related to dynamic scheduling, is the time at which the job is available for processing. Note that the static version of scheduling assumes that all jobs are simultaneously available for processing, i.e., zero ready times.

This paper considers the parallel machine scheduling problem in which each job may have sequence-dependent setup and distinct ready times. Note that these two characteristics make the problem more realistic, but increase its complexity drastically. The objective is to minimize total tardiness, i.e., amount of time that jobs fail to meet their due dates.

Compared with those for single machine scheduling, the research on parallel machine scheduling has been increased in last decades since various types of parallel machine systems are introduced to manufacturing and service systems. Since this paper focuses on the parallel machine scheduling problem with sequence-dependent setup and distinct ready times, the literature review is done on such problems. (See Cheng and Sin (1990) for literature reviews on various parallel machine scheduling problems.) Bean (1994) considers the parallel machine scheduling problem with sequence-dependent setup times for the objective of minimizing total tardiness, and suggests a genetic algorithm that can give near optimal solutions for the test problems, and later Koulamas (1997) suggests heuristic and simulated annealing algorithms for the same problem. Park and Kim (1997) consider the parallel machine scheduling problem with distinct ready times and suggest search heuristics that minimize work-in-process (WIP) inventory. For the problem with sequence-dependent setup and ready times, Sivrikaya-Serifoglu and Ulusoy (1999) suggests a genetic algorithm that minimizes the sum of the earliness and tardiness costs. Later, Bilge *et al.* (2004) suggest tabu search heuristics for the objective of minimizing total tardiness and show from computational tests that their algorithms outperform the genetic algorithm of Sivrikaya-Serifoglu and Ulusoy (1999) while the earliness cost is not considered. Recently, Cao *et al.* (2005) consider the parallel machine problem of simultaneously selecting and scheduling parallel machines, and suggest other tabu search heuristics for the objective of minimizing the sum of machine holding and job tardiness costs.

As stated earlier, this paper focuses on the parallel machine scheduling problem with sequence-dependent setup and ready times for the objective of minimizing total tardiness. In fact, the problem considered in this paper is the same as that of Bilge *et al.* (2004). Due to the complexity of the problem, we suggest tabu search heuristics that can give good solutions for practical sized problems within a reasonable amount of computation time. The algorithms suggested in this paper incorporate new methods to generate the neighborhood solutions. Three methods of obtaining the initial solutions are also suggested. Computational experiments are done on a number of randomly generated test problems and the results are reported. In particular, the new tabu search heuristics are compared with the existing one of Bilge *et al.* (2004).

2 Problem Description

Before describing the parallel machine scheduling problem considered in this paper, we present a general structure of parallel machine systems in Figure 1. In this figure, $j \parallel k$, implies job j is processed on machine k , $j = 1, 2, \dots, n$, $k = 1, 2, \dots, m$. Note that each job has a single operation that is performed on one of the parallel machines.

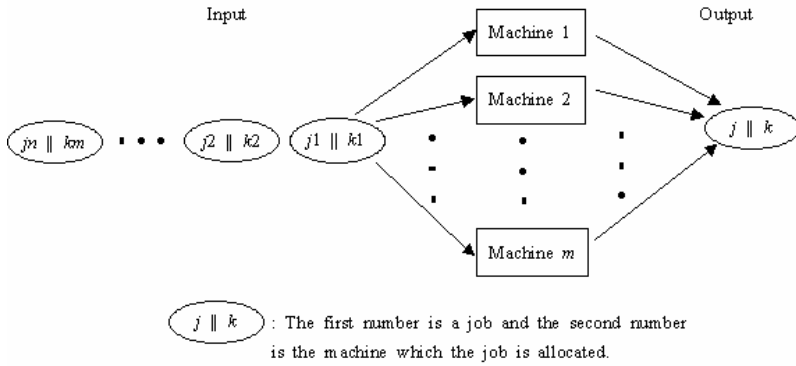


Fig. 1. Structure of parallel machine process

As stated earlier, there are two decision variables in the parallel machine scheduling problem: (a) allocating jobs to machines; and (b) sequencing the jobs assigned to each machine. In figure 1, the job allocation is denoted by $j \parallel k$ and the problems after allocating the jobs result in the single machine scheduling problems. It is assumed that the parallel machines are non-identical but uniform in that each machine is capable of processing all the jobs at different speeds. The objective is to minimize the total tardiness, which can be represented as

$$\sum_{i=1}^n \max\{0, C_i - d_i\},$$

where C_i and d_i denote the completion time and the due date of job i , respectively. Note that the completion times of jobs depend on the two decision variables, allocation and sequencing, and the problem considered here is to determine them while considering the sequence-dependent setup and ready times.

In this paper, we consider the deterministic version of the problem, i.e., processing time and due date of each job are deterministic and given in advance. It is assumed that sequence-dependent setup times are given. Hence, the time interval that job j occupies machine k can be represented as $s_{ijk} + p_{jk}$, where i is the job that precedes j in sequence on machine k , s_{ijk} is the setup time required for job j after job i is completed in machine k , and p_{jk} is the processing time of job j on machine k . It is also assumed that all jobs are known at the beginning of the scheduling horizon, and their ready times are distinct and given in advance. Recall that the ready time of a job is the time at which the job is available for processing.

Other assumptions made in the problem considered here are the same of those of the basic single machine scheduling problem. They are: (a) jobs are independent, i.e., no precedence relationship between any pair of jobs; (b) each machine can process only one job at a time and each job can be processed on one machine; (c) once a job is determined to be processed by a machine, it will stay on the machine until its completion, i.e., no job preemption; and (d) machine breakdowns are not considered. As noted in the previous research articles, the problem considered in this paper is an NP-hard problem. This can be easily seen from the fact that the single machine problem with the tardiness measure is NP-hard even without sequence-dependent setup and ready times (Du and Leung 1990).

3 Solution Algorithms

3.1 Obtaining Initial Solutions

We suggest three methods to obtain the initial solutions, each of which is used for the tabu search heuristics suggested in this paper. Detailed explanations of the three methods are given below.

EDR (Earliest Due date plus Ready time)

This method is an extension of the EDD (Earliest Due Date) rule, a commonly used one for various scheduling problems with due dates. Because we consider the problem with ready times, EDR uses the sum of due date and ready time as the sorting measure. More specifically, all jobs are sorted in a nondecreasing order of these sums, and the jobs are allocated to machines in this order. Also, the job allocation is done by selecting the machine with the smallest additional time. Here, the additional time of job j on machine k is defined as $p_{jk} - p_{jk_i}$, where p_{jk} is the processing time of job j in machine k (as defined earlier) and k_j is the index for the machine where the smallest processing time of job j occurs. For example, consider two types (old and new) of machines. If the processing times of a job are 3 and 4 at the new and the old machines, respectively, the additional time of the old (new) one becomes 1 (0).

MT (Minimum Tardiness)

This is similar to the EDR method in that the job sorting measure is the sum of due date and ready time, and all jobs are sorted in a nondecreasing order of these sums. Unlike the EDR method, however, the job allocation is done by selecting the machine with the smallest total tardiness after the current job is allocated. (Note that the total tardiness is calculated for each machine.) In this method, tie breaks are done as follows. If there are two or more machines with zero or the same total tardiness, the machine with the maximum sum of slack time is selected. Here, the slack time of a job is the difference between its due date and completion time.

MPS (Minimum Partial Schedule)

This is the same as the MT method except that a set of tardy jobs is maintained during the job allocation. That is, if a job can be allocated to one or more machines without incurring tardiness, selected is the machine with the smallest completion time after the job is allocated. Otherwise, the job (that incurs tardiness) is assigned to the set of tardy jobs. Basically, this idea is similar to the Hodgson algorithm for the basic single machine scheduling problem that minimizes the number of tardy jobs. After all jobs are considered for allocation, a partial schedule can be obtained for the allocated job, and then the tardy jobs are scheduled. To schedule each of the tardy jobs, the best position (that gives the minimum total tardiness) is selected after it is assigned to all possible positions in the current partial schedule.

3.2 Tabu Search Heuristics

Tabu search (TS) is one of the well-known local search techniques that have been successfully applied to various combinatorial optimization problems. Starting from an

initial solution, TS generates a new alternative S' in the neighborhood of the original alternative S with a function that transforms S into S' . This is usually called a *move*, which can be made to a neighbor solution even though it is worse than the given solution. This makes a TS escape from a local optimum in its search for the global optimum. To avoid cycling, TS defines a set of moves that are tabu (forbidden), and these moves are stored in a set A , called *tabu list*. Elements of A define all tabu moves that cannot be applied to the current solution. The size of A is bounded by a parameter l , called *tabu list size*. If $|A| = l$, before adding a move to A , one must remove an element in it, the oldest one in general. Note that a tabu move can be always allowed to be chosen if it creates a solution better than the incumbent solution, the best objective value obtained so far. This is called the *aspiration criterion*. See Glover (1989) for a comprehensive description of various aspects of TS.

An application of TS is generally characterized by several factors. They are: (a) initial solution methods; (b) neighborhood generation methods, i.e., set of possible moves applicable to the current solution; (c) definition of tabu moves with the tabu list size; and (d) termination condition(s).

First, the solution is represented as Figure 2, which is modified from that of Bilge *et al.* (2004). In this figure, J_{kd} denotes the index of the d th job assigned to machine k , and n_k denotes the number of jobs allocated to machine k . Also, the objective value for a given solution can be easily calculated by going through the sequence of jobs over each machine and summing up the tardiness of each job, calculated by taking into account distinct ready times, due dates, processing times on different types of machines, and sequence-dependent setup times.

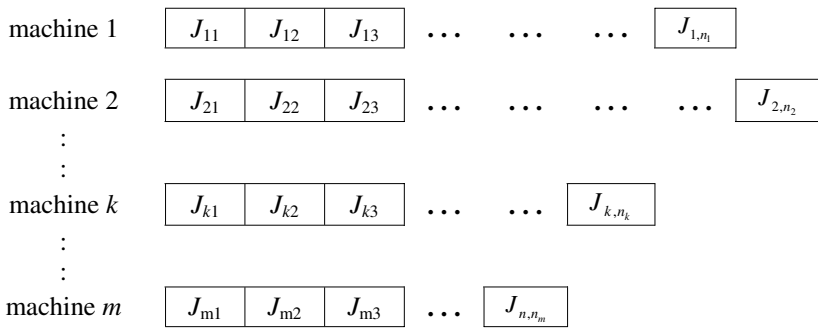


Fig. 2. Solution representation (adopted from Bilge *et al.* 2004)

Two neighborhood generation methods are suggested in this paper, each of which is based on swap and insertion methods. In general, the swap method, alternatively called the interchange method in the literature, generates a neighborhood solution by selecting two jobs in the current schedule and interchanging them, while the insertion method generates a neighborhood solution by randomly selecting two jobs and removing the first job from the original place and inserting it to the position that directly precedes the second job. Note that the swap and insertion methods can be done for the jobs assigned to the same machine (intra-machine) or different machines (inter-machine) since this paper considers the parallel machine scheduling problem.

The neighborhood generation methods suggested in this paper have a hybrid structure with swap and insertion methods. A schematic description of the hybrid structure is given in Figure 3. As can be seen from the figure, the hybrid structure uses the swap and the insertion methods consecutively. More specifically, the swap move is done first for a pair of jobs, and then the insertion move is done for each of the neighborhoods generated by the swap move. Here, the intra- and inter-machine moves are considered for both swap and insertion methods. Note that Bilge *et al.* (2004) considers the intra- and inter-machine moves for the insertion method, but only the intra-machine moves for the swap method. Therefore, our neighborhood generation methods extend the method of Bilge *et al.* (2004) in that the search space is enlarged and hence the possibility of obtaining good solutions is increased.

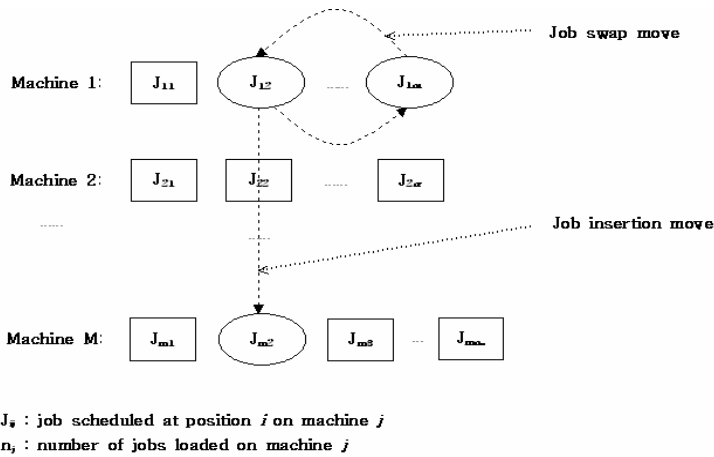


Fig. 3. A schematic description of neighborhood generation

There may be several ways of considering moves. Among them, this paper uses the method of examining a portion of the entire neighborhood and taking the best move that is not tabu. This is because the hybrid method described earlier generates too many neighborhood solutions, and hence it is needed to limit the number of neighborhood solutions examined. This is called the *candidate list strategy* in the literature (Laguna *et al.* 1991). The purpose of the candidate list strategy is to screen the neighborhood solutions so as to concentrate on promising moves. This paper suggests two candidate list strategies, called *small candidate list strategy* (SCLS) and *large candidate list strategy* (LCLS) in this paper. Note that each of the two strategies is applied to the hybrid neighborhood generation method. In the SCLS, only one job with the maximum tardiness is selected for each machine, and these jobs are considered as the candidates for (intra-machine and inter-machine) insertion or swap move. Since this approach has shown to be quite fast and decreases the neighborhood size considerably, it is called as the small candidate list strategy. In the LCLS, on the other hand, two cases are considered. If the iteration number is odd, all tardy jobs on each machine are considered as the candidates for insertion or swap move. Otherwise, all non-tardy jobs on each machine are chosen as the candidates. Here, the insertion

and swap moves are done for both intra-machine and inter-machine. In this paper, therefore, we suggest six TS heuristics, each of which is a combination of the initial solution methods described in section 3.1 and the two neighborhood generation methods.

Tabu moves are defined as follows. In the swap move, a pair of jobs that have been interchanged is defined as a tabu move. Also, the insertion method defines a tabu move as the job to be moved and the job that directly precedes the second job. More specifically, if job i is inserted between jobs $(j - 1)$ and j , jobs i and j are stored in the tabu list. Here, in the case that job i is inserted to the last position in the sequence of jobs assigned to a machine, job i and the last job in the sequence are defined as a tabu move. As an exceptional case, a tabu move can be allowed to be chosen if it generates a solution better than the incumbent solution, the best objective value obtained so far.

As stated earlier, the size of the tabu list is an important parameter. If the tabu list is too short, the TS heuristics may keep returning to the same local optimum, which prevents the search process from exploring a wide area of solution space. In contrast, if the tabu list is too long, it results in excessive computation time searching the tabu list to determine if a move is tabu or not. Thus, a longer time spent going through a tabu list provides less time for the procedure to explore the solution space for a given computation time. In this paper, the tabu list size l was determined using the empirical function, suggested by of Bilge *et al.* (2004), which depends on the problem size. Formally, it is represented as

$$l = h \cdot \frac{n \cdot \sqrt{n}}{m - 0.5},$$

where h is a parameter. Also, two independent tabu lists, which correspond to the swap and the insertion methods for generating neighborhoods, are created, and each of them is maintained in such a way that the oldest tabu move is removed before adding a new one if its tabu list is full.

Finally, the TS heuristics stop if no improvements have been made for a certain number of consecutive iterations, denoted by L_{TS} in this paper. This is a common termination rule used in the literature.

4 Computational Experiments

To show the performances of the TS heuristics, computational tests were done on randomly generated test problems, and the results are reported in this section. In the test, 7 TS heuristics are compared, one suggested by Bilge *et al.* (2004) and six suggested in this paper. The TS heuristics were coded in C, and tests were performed on a workstation with an Intel Xeon processor operating at 3.2 GHz clock speed.

Since optimal solutions cannot be obtained in a reasonable amount of computation time, the TS heuristics are compared by using a relative performance ratio. Here, the relative performance ratio of the TS heuristic a for a problem is defined as

$$100 \cdot (C_a - C_{best}) / C_{best},$$

where C_a is the objective value obtained using the heuristic a for the problem and C_{best} is the best objective function value among those obtained from the seven TS heuristics.

Table 1. Performances of the TS heuristics

(a) Cases of loose due dates

Machines	Jobs	Bilge et al. ¹	EDR-SCLS ²	MT-SCLS ²	MPS-SCLS ²	EDR-LCLS ²	MT-LCLS ²	MPS-LCLS ²
4	20	1.4(4.3)*	0.5(1.0)	0.2(0.3)	0.1(0.3)	0.0(0.0)	0.0(0.0)	0.0(0.0)
	40	0.5(0.9)	3.3(3.6)	2.0(2.6)	2.6(3.1)	0.1(0.3)	0.2(0.1)	0.1(0.3)
	60	0.5(1.4)	4.6(3.5)	6.0(4.4)	5.7(4.2)	2.0(3.6)	1.9(3.6)	2.9(4.0)
	80	0.1(0.2)	2.4(3.1)	1.4(1.8)	1.3(1.6)	0.3(1.0)	0.3(1.0)	0.3(1.0)
8	60	1.2(1.8)	2.1(1.8)	2.4(2.7)	3.3(2.8)	0.0(0.0)	0.0(0.0)	0.0(0.1)
	80	1.6(2.6)	3.7(2.8)	3.8(2.7)	3.1(3.3)	0.4(1.1)	0.3(0.7)	0.4(1.1)
	100	0.8(1.3)	6.2(6.9)	6.9(8.0)	10.7(13.6)	1.2(2.6)	0.8(1.7)	0.4(1.1)
	120	0.2(0.3)	3.1(3.2)	3.1(3.2)	4.2(5.9)	0.3(0.9)	0.3(0.9)	0.3(0.9)
12	100	1.1(1.0)	3.7(2.1)	4.2(2.1)	4.0(1.7)	0.5(0.8)	0.4(0.5)	0.4(0.8)
	120	1.4(2.4)	4.0(4.8)	4.5(3.9)	4.5(3.3)	0.4(0.9)	0.4(0.8)	0.6(0.9)
	140	0.4(1.5)	5.0(2.3)	4.6(2.9)	4.1(2.0)	0.6(0.8)	0.8(0.8)	0.7(0.8)
	160	1.6(3.3)	4.8(1.9)	5.9(2.0)	5.1(2.0)	0.4(0.8)	0.7(1.0)	0.6(1.0)

¹ Algorithm by Bilge et al. (2004)

² EDR, MT, MPS: Initial solutions SCLS, LCLS: Candidate selection strategies of neighborhood solutions

* Average value of RPRs for 10 cases (standard deviation in parentheses)

(b) Cases of medium due dates

Machines	Jobs	Bilge et al. ¹	EDR-SCLS ²	MT-SCLS ²	MPS-SCLS ²	EDR-LCLS ²	MT-LCLS ²	MPS-LCLS ²
4	20	3.6(4.0)	3.6(2.7)	4.1(3.2)	2.8(2.8)	0.1(0.1)	0.0(0.1)	0.0(0.0)
	40	9.3(6.5)	20.0(16.4)	21.2(15.4)	22.5(17.6)	0.5(0.9)	0.7(1.0)	0.8(0.7)
	60	9.2(17.9)	14.5(10.2)	16.0(16.5)	20.6(18.0)	1.1(2.6)	1.2(2.0)	0.4(0.6)
	80	13.8(28.2)	15.9(17.9)	16.8(21.9)	14.0(16.4)	1.1(2.9)	2.1(3.3)	1.3(3.0)
8	60	4.0(2.0)	14.0(5.7)	13.8(5.5)	14.4(9.1)	0.8(1.1)	0.7(1.2)	1.1(1.8)
	80	3.5(3.2)	8.8(4.4)	9.3(4.4)	9.6(5.5)	0.2(0.3)	0.4(0.7)	0.4(0.7)
	100	4.4(4.4)	13.7(6.0)	12.9(6.2)	14.4(7.1)	0.3(0.5)	0.3(0.4)	0.4(0.5)
	120	10.2(25.6)	15.9(10.0)	10.6(6.2)	14.7(9.7)	2.3(6.7)	2.3(6.5)	2.5(6.7)
12	100	9.0(13.0)	21.7(23.2)	21.0(24.1)	20.9(23.5)	0.0(0.1)	3.8(10.8)	4.5(11.8)
	120	3.6(4.5)	9.7(7.3)	9.0(5.4)	9.1(6.4)	0.2(0.2)	0.4(0.5)	0.3(0.5)
	140	1.2(1.6)	7.2(2.7)	7.6(3.3)	7.5(2.8)	0.2(0.3)	0.3(0.3)	0.2(0.3)
	160	1.7(2.0)	8.6(4.0)	8.6(4.2)	9.0(3.5)	0.6(0.7)	0.4(0.4)	0.5(0.4)

(c) Cases for tight due dates

Machines	Jobs	Bilge et al. ¹	EDR-SCLS ²	MT-SCLS ²	MPS-SCLS ²	EDR-LCLS ²	MT-LCLS ²	MPS-LCLS ²
4	20	1.4(1.5)	1.1(1.2)	1.4(1.2)	0.9(1.0)	0.1(0.1)	0.1(0.3)	0.1(0.2)
	40	7.5(4.3)	12.4(4.2)	13.6(4.2)	14.7(6.6)	0.2(0.4)	0.3(0.5)	0.5(0.7)
	60	4.8(3.7)	28.5(16.7)	29.1(16.7)	26.3(14.1)	0.1(0.2)	2.3(3.4)	1.3(2.1)
	80	8.8(11.7)	29.5(15.4)	31.6(15.4)	26.6(12.7)	0.5(0.8)	0.5(1.4)	0.7(1.1)
8	60	3.2(1.7)	13.5(4.7)	12.4(4.7)	12.5(4.6)	1.2(1.0)	0.1(0.2)	1.3(1.0)
	80	3.8(1.8)	21.0(5.8)	21.5(5.8)	22.2(5.7)	0.5(0.6)	0.7(0.9)	0.7(1.1)
	100	6.9(6.5)	23.2(6.9)	21.6(6.9)	25.3(10.7)	0.7(1.2)	1.4(1.8)	0.5(0.7)
	120	8.2(13.1)	25.8(14.1)	26.1(14.1)	27.8(18.8)	0.8(1.1)	0.5(0.9)	2.0(3.4)
12	100	7.0(4.9)	18.2(5.8)	16.6(5.8)	17.1(3.9)	1.4(1.5)	1.1(1.3)	1.4(1.6)
	120	4.6(3.0)	16.4(7.0)	14.5(7.0)	15.4(5.2)	1.3(2.2)	0.5(0.8)	1.0(1.7)
	140	14.6(35.8)	25.1(8.7)	25.3(8.7)	25.4(10.6)	1.6(1.8)	1.7(3.3)	0.5(0.7)
	160	5.0(4.5)	19.6(18.5)	16.6(18.5)	18.8(16.5)	0.9(1.2)	0.5(0.7)	1.2(3.1)

For the test, 360 problems were generated randomly, i.e., 10 problems for 36 combinations of three levels of the number of machines (4, 8 and 12), eight levels of the number of jobs (20, 40, 60, 80, 100, 120, 140 and 160), and three levels of the due date tightness (loose, medium, tight). More specifically, the number of jobs was set to 20, 40, 60 and 80 for the case of 4 machines, 60, 80, 100 and 120 for the case of 8 machines, and 100, 120, 140 and 160 for the case of 12 machines. The test problems were generated using the method of Sivrikaya-Serifoglu and Ulusoy (1999), which was also adopted by Bilge *et al.* (2004). Details of the problem generation method are omitted here because of the space limitation. Also, to find the most appropriate parameter values for the TS heuristics, several values for the parameters for the tabu list size (h) and the termination condition (L_{TS}) were tested on several representative test problems, and they were set to 1 and 5000, respectively.

Results for the test problems are summarized in Table 1 which shows the relative performance ratios of the seven TS heuristics for the cases of loose, medium, and tight due dates, respectively. It can be seen from the table that the TS heuristics with the LCLS strategy outperform the others. Although it is not reported, statistical tests also showed statistically significant difference between the heuristics with the LCLS strategy and the others. Also, no significant differences could be found among the three TS heuristics with the LCLS strategy. This implies the performances of the TS heuristics suggested in this paper do not depend on the initial solution method. In particular, the three TS heuristics outperform the existing algorithm of Bilge *et al.* (2004) significantly. More specifically, we could obtain improvements from the three TS heuristics about 1%, 6%, and 8% in overall average for the cases of loose, medium and tight due dates (Table 1(a), (b), and (c)), respectively. Also, due to their large search space, the heuristics with the LCLS strategy required longer computation times than those with the SCLS strategy. However, most of the test problems were solved within 1 hour even for the large-sized test problems with 12 machines and 160 jobs, which is reasonable for practical scheduling problems, e.g., an injection molding shops from our experience.

5 Concluding Remarks

This paper considered the parallel machine scheduling problem with sequence-dependent setup and ready times for the objective of minimizing total tardiness. The machines are uniform in that they can process all the jobs with difference processing and setup rates. Due to the complexity of the problem, we suggested tabu search heuristics together with three initial solution methods (EDR, MT, and MPS) and two neighborhood generation methods (SCLS and LCLS). To show the performances of the heuristics, computational experiments on randomly generated test problems were done and the results showed that the tabu search heuristics with LCLS outperformed the others due to their extended search space. In particular, they gave significant improvements over an existing heuristic.

This research can be extended as follows. First, in the theoretical aspect, it may be needed to develop optimal algorithms. To the best of our knowledge, there is no other research article that suggests an optimal algorithm for generalized parallel machine scheduling with sequence-dependent setup and ready times. Second, more sophisticated neighborhood generation methods such as the intensification and

diversification strategies can be developed for the search heuristics. Finally, other search techniques can also be applied to this complex problem.

Acknowledgements

This work was supported by Korea Research Foundation Grant funded by Korean Government (MOEHRD) (KRF-2005-041-D00893). This is gratefully acknowledged.

References

1. Bean, J. C.: Genetic algorithm and random keys for sequencing and optimization, *ORSA Journal on Computing*, Vol. 6 (1994) 154-160.
2. Bilge, U., Kirac, F., Kurtulan, M., and Pekgun, P.: A tabu search algorithm for parallel machine total tardiness problem, *Computers and Operations Research*, Vol. 31 (2004) 397-414.
3. Cao, D., Chen, M., and Wan, G.: Parallel machine selection and job scheduling to minimize machine cost and job tardiness, *Computers and Operations Research*, Vol. 32 (2005) 1995-2012.
4. Cheng, T. C. E. and Sin, C. C. S.: A state-of-the-art review of parallel-machine scheduling research, *European Journal of Operational Research*, Vol. 47 (1990) 271-292.
5. Dastidar, G. S. and Nagi, R.: Scheduling injection molding operations with multiple resource constraints and sequence dependent setup times and costs, *Computers and Operations Research*, Vol. 32 (2004) 2987-3005.
6. Du, J. and Leung, J. Y. T.: Minimizing total tardiness on one machine is NP-hard, *Mathematics of Operations Research*, Vol. 15 (1990) 483-495.
7. Glover, F.: Tabu Search: Part I, *ORSA Journal on Computing*, Vol. 1 (1989) 190-206.
8. Koulamas, C.: Decomposition and hybrid simulated annealing heuristics for the parallel machine total tardiness problem, *Naval Research Logistics*, Vol. 44 (1997) 109-125.
9. Laguna, M., Barnes, J. W., and Glover, F.: Tabu search methods for a single machine scheduling problem, *Journal of Intelligent Manufacturing*, Vol. 2 (1991) 63-74.
10. Park, M.-W. and Kim, Y.-D.: Search heuristics for a parallel machine scheduling problem with ready times and due dates, *Computers and Industrial Engineering*, Vol. 33 (1997) 793-796.
11. Sivrikaya-Serifoglu, F. and Ulusoy, G.: Parallel machine scheduling with earliness and tardiness penalties, *Computers and Operations Research*, Vol. 26 (1999) 773-787.

The Maximum Integer Multiterminal Flow Problem

Cédric Bentz

CEDRIC-CNAM, 292, Rue Saint-Martin,
75141 Paris Cedex 03, France
cedric.bentz@cnam.fr

Abstract. Given an edge-capacitated graph and k terminal vertices, the *maximum integer multiterminal flow* problem (MAXIMTF) is to route the maximum number of flow units between the terminals. For directed graphs, we introduce a new parameter $k_L \leq k$ and prove that MAXIMTF is \mathcal{NP} -hard when $k = k_L = 2$ and when $k_L = 1$ and $k = 3$, and polynomial-time solvable when $k_L = 0$ and when $k_L = 1$ and $k = 2$. We also give an $2 \log_2(k_L + 2)$ -approximation algorithm for the general case. For undirected graphs, we give a family of valid inequalities for MAXIMTF that has several interesting consequences, and show a correspondence with valid inequalities known for MAXIMTF and for the associated *minimum multiterminal cut problem*.

1 Introduction

Routing problems in networks are commonly modeled by flow or multicommodity flow problems. Given an edge-capacitated graph (directed or undirected), the goal is to route flow units (requests) between prespecified vertices. When one seeks to route the maximum number of flow units from a unique source to a unique sink, the problem is the famous *maximum flow problem*. The Ford-Fulkerson's theorem [11] gives a good characterization for this case, which is efficiently solvable [1]. In particular, this theorem states that, if the capacities are integral, the value of a maximum integer flow is equal to the value of a minimum cut, i.e., to the value of a minimum weight set of edges whose removal separates the source from the sink. Unfortunately, this does not hold for more general variants. One of the most studied variant is the *maximum integer multi-commodity flow problem*: given an edge-capacitated graph $G = (V, E)$ and a list of source-sink pairs, the goal is to simultaneously route the maximum number of flow units, each unit being routed from one source to its corresponding sink.

This problem is \mathcal{NP} -hard even for two source-sink pairs [10], and cannot be approximated within $|E|^{1/2-\epsilon}$ (resp. within $(\log |E|)^{1/3-\epsilon}$) for every $\epsilon > 0$ in directed graphs [15] (resp. in undirected graphs [2]) unless $\mathcal{P} = \mathcal{NP}$ (recall that, for a maximization (resp. minimization) problem, an α -approximation algorithm is a polynomial-time algorithm that always outputs a feasible solution whose value is at least $1/\alpha$ times (resp. at most α times) the value of an optimal solution). The corresponding generalization of the problem of finding a minimum cut is the

minimum multicut problem, which asks to select a minimum weight set of edges whose removal separates each source from its corresponding sink. This problem is also \mathcal{NP} -hard in several special cases, and has a noticeable relationship with the former: the continuous relaxations of the linear programming formulations of the two problems are dual [8]. In particular, this interesting property has been used to design good approximation algorithms for both problems [14]. Further results and references concerning these problems can be found in [1] and [8].

Another generalization of the maximum flow problem is the *maximum integer multiterminal flow problem* (MAXIMTF): given an edge-capacitated graph and a set $T = \{t_1, \dots, t_k\}$ of *terminal* vertices, MAXIMTF is to route the maximum number of flow units between the terminals. Note that this problem is a particular maximum integer multicommodity flow problem in which the source-sink pairs are (t_i, t_j) for $i \neq j$. The associated *minimum multiterminal cut problem* (MINMTC) is to select a minimum weight set of edges whose removal separates t_i from t_j for $i \neq j$. Note that MAXIMTF and MINMTC also have the duality relationship mentioned above. MINMTC has been widely studied in the undirected case (see [3], [5], [6], [7], [8], [9], [16] and [20]), and the directed case has also received some attention: Garg et al. [13] show that it is \mathcal{NP} -hard even for $k = 2$ and give an $2 \log_2 k$ -approximation algorithm, and Naor and Zosin [18] give a 2-approximation algorithm. However, the algorithm of Garg et al. has an interesting property: it computes a multiterminal cut whose value is at most $2 \log_2 k$ times the value of an integer multiterminal flow, and hence is an $2 \log_2 k$ -approximation for both MINMTC and MAXIMTF (while the algorithm of Naor and Zosin does not provide an approximate solution for MAXIMTF). (Note that the same idea easily yields an $\log_2 k$ -approximation algorithm for MAXIMTF in undirected graphs.) Costa et al. [8] show that MAXIMTF and MINMTC are polynomial-time solvable in acyclic directed graphs by using a simple reduction to a maximum flow and a minimum cut problem, respectively. To the best of our knowledge, these are the only results about MAXIMTF in directed graphs. In undirected graphs, MAXIMTF has recently be shown to be polynomial-time solvable by using the ellipsoid method [17] (the result is based on the associated Mader's theorem on T -paths [19, Chap. 73]). Algorithmic aspects of special cases have also been studied (inner eulerian graphs in [12] and trees in [4]). However, it can be easily noticed that, for all the problems mentioned above, the general directed case is "harder" than the undirected one, since there exists a linear reduction from the latter to the former: simply replace each edge by the gadget given in [19, (70.9) on p. 1224].

The motivation of this paper is to explore further the complexity of MAXIMTF. Given a directed graph, we say that a terminal is *lonely* if it lies on at least one directed cycle containing no other terminal, and we let T_L denote the set of lonely terminals and $k_L = |T_L|$. We shall see that k_L is a key parameter for better understanding the complexity and approximability of MAXIMTF. Moreover, some of our results will extend to MINMTC.

We first show that MAXIMTF is strongly \mathcal{NP} -hard in directed graphs, even if $k_L = k = 2$ or if $k_L = 1$ and $k = 3$ (Section 2). Then, we prove MAXIMTF to

be tractable when $k_L = 0$ and when $k_L = 1$ and $k = 2$, and improve the $2 \log_2 k$ -approximation algorithm of Garg et al. [13] by providing an $2 \log_2(k_L + 2)$ -approximation algorithm for the general case (Section 3). Eventually, we give a family of valid inequalities for MAXIMTF in undirected graphs, and show an interesting correspondence with valid inequalities known for the associated problem MINMTC (Section 4).

Note that, throughout this paper, we consider only *simple* graphs. We call *Directed* (resp. *Undirected*) MAXIMTF the problem MAXIMTF defined in directed (resp. undirected) graphs. Moreover, due to lack of space, we sometimes omit some details in our proofs.

2 \mathcal{NP} -Hardness Proof

We show in this section that Directed MAXIMTF is strongly \mathcal{NP} -hard, even if $k = k_L = 2$ (or $k_L = 1$ and $k = 3$). In order to do this, we adapt the proof, given in [10], of the \mathcal{NP} -completeness of the *directed integer multicommodity flow problem* with two source-sink pairs, (s_1, s'_1) and (s_2, s'_2) : given an arc-capacitated directed graph $G = (V, A)$ and two integer demands d_1 and d_2 associated with the respective source-sink pairs, it asks to decide whether these demands can be simultaneously routed while respecting the capacity constraints (if, for an instance, the answer is *yes*, then this instance is *solvable*). In the instance used in the proof of [10, Theorem 3], $d_1 = 1$, $d_2 \leq |V|$ and all the arcs have capacity 1. Moreover, this instance satisfies

$$|\Gamma^-(s_1)| = |\Gamma^-(s_2)| = 0$$

and

$$|\Gamma^+(s'_1)| = |\Gamma^+(s'_2)| = 0$$

where, for $v \in V$, $\Gamma^+(v) = \{u \in V \text{ such that } (v, u) \in A\}$ and $\Gamma^-(v) = \{u \in V \text{ such that } (u, v) \in A\}$. We modify this initial instance as follows: we add two new vertices, t_1 and t_2 , and four arcs (t_1, s_1) , (s'_2, t_1) , (t_2, s_2) and (s'_1, t_2) , valued by 1, d_2 , d_2 and 1 respectively (see Fig. 1(a)).

It is easy to see that the initial instance is solvable if and only if the optimum value for the maximum integer multicommodity flow instance defined on the pairs (t_1, t_2) and (t_2, t_1) is equal to $d_2 + 1$ (no flow unit being routed from s_i to s_j , from s'_i to s'_j for $i \neq j$, or from s_i to s'_j for $i \neq j$). Moreover, the latter instance is equivalent to a directed maximum integer multiterminal flow instance with two terminals, t_1 and t_2 . Eventually, we can replace each one of the two arcs (s'_2, t_1) and (t_2, s_2) by d_2 directed paths of length two (containing only arcs with capacity 1) between the corresponding endpoints, and obtain an instance of MAXIMTF where each arc has capacity 1. The described transformation is clearly polynomial, and hence

Theorem 1. *Directed MAXIMTF is \mathcal{NP} -hard in graphs with unit capacities, even with only two terminals.*

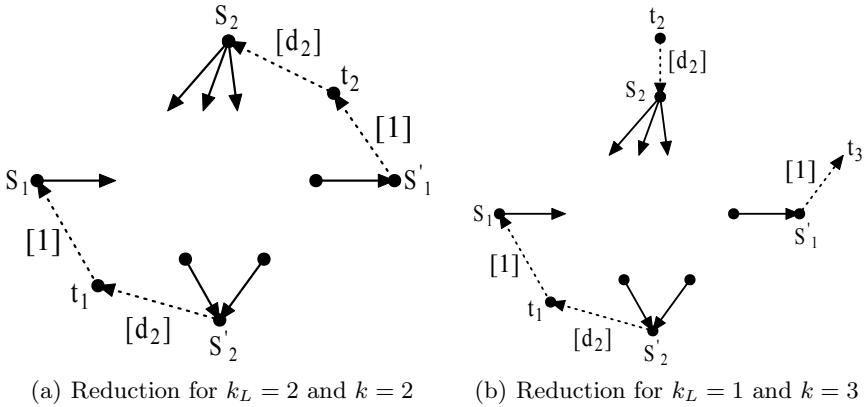


Fig. 1. Reductions for Directed MAXIMTF

In particular, this implies the strong \mathcal{NP} -hardness of Directed MAXIMTF. It can also be noticed that this result matches the complexity result for the associated cut problem MINMTC in directed graphs [13].

However, in the proof of Theorem 1, $k = k_L = 2$: so, what happens when $k_L = 1$? Actually, a slightly different proof shows that Directed MAXIMTF remains \mathcal{NP} -hard, even with unit capacities. We define three new terminals instead of two: t_1, t_2 and t_3 . Moreover, we add four arcs $(t_1, s_1), (s'_2, t_1), (t_2, s_2)$ and (s'_1, t_3) , valued by 1, d_2, d_2 and 1 respectively (note that $T_L = \{t_1\}$; see Fig. 1(b)). In this instance, it is easy to see that there exists an integer multiterminal flow of value $d_2 + 1$ if and only if 1 flow unit is routed from t_1 to t_3 and d_2 flow units are routed from t_2 to t_1 . This implies:

Theorem 2. Directed MAXIMTF is \mathcal{NP} -hard in graphs with unit capacities, even if $k_L = 1$ and $k = 3$.

We shall deal with the cases where $k_L = 0$ and where $k_L = 1$ and $k = 2$ in the next section.

3 Exact and Approximation Algorithms

From the previous section, Directed MAXIMTF is strongly \mathcal{NP} -hard even for $k_L = 1$ and $k = 3$ and for $k = k_L = 2$. Hence, if $\mathcal{P} \neq \mathcal{NP}$, the only efficient algorithms one can expect to design are approximation algorithms. In this section, we improve the $2 \log_2 k$ -approximation algorithm of Garg et al. [13] and give an $2 \log_2(k_L + 2)$ -approximation algorithm for Directed MAXIMTF.

The basic idea of our approach is to combine the algorithm of Garg et al. with an interesting strengthening of [8, Proposition 3]. The main idea of the proof of [8, Proposition 3] (that shows that MAXIMTF and MINMTC are polynomial-time solvable in acyclic directed graphs) is to split up each terminal vertex t_i

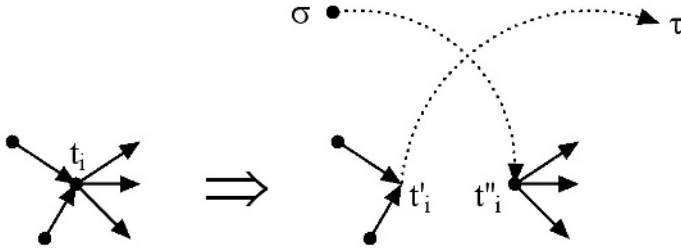


Fig. 2. Splitting up terminal t_i

into two new vertices, t'_i and t''_i , such that all the vertices in $\Gamma^-(t_i)$ are linked to t'_i and t''_i is linked only to the vertices in $\Gamma^+(t_i)$. Then, we add two new vertices, σ and τ , and link (by arcs with sufficiently large capacities) every t'_i to τ and σ to every t''_i (see Fig. 2). Finally, we compute a maximum flow between σ and τ (obviously, we assume that the capacities are integral). The obtained flow is a valid integer multiterminal flow for the initial instance if, in the modified instance, no flow unit is routed from t''_i to t'_i for some i .

The main point for us is that, if there is no lonely terminal, then, by splitting up the terminals as explained, there will remain no directed path from t''_i to t'_i for each i , and hence we will be able to solve MAXIMTF and MINMTC using the above technique. Actually, if we want to guarantee that, after splitting up each terminal, the modified graph does not admit a directed path from t''_i to t'_i for some i (otherwise, we cannot be sure that the flow we will compute in the modified graph will be a valid multiterminal flow in the initial graph), this is essentially the best (i.e., weakest) assumption that can be made. Namely, one can easily show:

Theorem 3. *After splitting up all the terminals, there is no directed path between t''_i and t'_i for each i if and only if $k_L = 0$.*

This also implies the following strengthening of [8, Proposition 3]:

Theorem 4. *MINMTC and MAXIMTF are polynomial-time solvable in directed graphs if $k_L = 0$, by using a max flow-min cut algorithm.*

Actually, the last remaining case, i.e., the case where $k_L = 1$ and $k = 2$, is also polynomial-time solvable. Indeed, one can prove that on any directed cycle containing only the terminal in T_L , there is a *removable* arc, i.e., an arc lying on no elementary path between the two terminals (otherwise, there would be a vertex of this cycle lying on another directed cycle containing only the terminal not in T_L , which is impossible). By iteratively removing such arcs, we are back to the case where $k_L = 0$. Hence:

Theorem 5. *Directed MAXIMTF is tractable if $k_L = 1$ and $k = 2$.*

Theorems 3 and 4 show the importance of the parameter k_L for both MAXIMTF and MINMTC. Moreover, this suggests the following approach for finding

approximate solutions for these two problems: first, (a) split up each terminal $t_i \in T - T_L$ into t'_i and t''_i as explained above, add the two vertices σ and τ , and link (by *heavy arcs*) every t'_i to τ and σ to every t''_i ; then, (b) compute a multiterminal cut and flow for this new instance (i.e., where the terminal set is $T_L \cup \{\sigma, \tau\}$) by using the algorithm of Garg et al. [13]. The definition of T_L guarantees that we obtain a valid integer multiterminal flow.

Hence, the main difference with their algorithm is that, before using their divide-and-conquer strategy, we transform the graph by replacing the terminals in $T - T_L$ by two new terminals, σ and τ . This implies that we use Garg et al.'s algorithm on an instance with $k_L + 2$ terminals, and so we obtain an approximation factor of $2 \log_2(k_L + 2)$ (instead of $2 \log_2 k$).

Actually, one can even prove that this analysis of the approximation ratio is tight by using a particular family of instances built on an undirected tree with $k = 2^p$ vertices (all vertices are terminal), and which is transformed into a directed graph by replacing each edge by the gadget given in [19, (70.9) on p. 1224] (each arc of the gadget has the capacity of the initial edge). Due to lack of space, we do not give the whole construction here.

4 Polyhedral Results for the Undirected Case

In this section, we give a family of valid inequalities for the LP formulation of Undirected MAXIMTF given in [13]. We call them *tree inequalities*, as they can be seen as the “flow counterpart” of the tree inequalities given in [7] and characterizing completely the polytope of MINMTC in trees (as shown in [7]).

Theorem 6. *Let U be an undirected tree and $X_U = \{t_1, \dots, t_h\}$ be the terminals in U . Assume the leaves of U coincide with these h terminals (assume without loss of generality that we have removed from U the edges such that no flow unit can be routed through them). Then, if F_U denotes the total number of flow units that are routed between the terminals in X_U , the inequality*

$$F_U \leq \left\lfloor \frac{\sum_{i \in \{1, \dots, h\}} c_i}{2} \right\rfloor$$

is a valid inequality for MAXIMTF (called tree inequality), where, for each i , c_i is the minimum capacity of the edges contained in the path p_i linking t_i to n_i , its nearest vertex of degree at least 3 in U (see Fig. 3).

A proof of this theorem is given below. In particular, this proof will imply that these inequalities (together with the usual constraints) are sufficient to guarantee the existence of an integer optimal solution to the continuous relaxation of the linear program (i.e., of an optimal solution for MAXIMTF) in undirected trees (recall that the tree inequalities for MINMTC do give a complete characterization of the associated polytope in undirected trees [7]). These inequalities may also be sufficient to completely characterize the polytope associated with MAXIMTF in undirected trees. On the other hand, a quadratic algorithm for

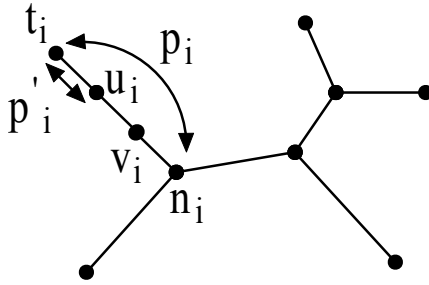


Fig. 3. A tree instance for Theorem 6 ((u_i, v_i) has capacity c_i)

MAXIMTF in trees is already known [4]. Our original motivation came from the fact that the authors of [7] used the description of the polytope associated with MINMTC to derive an efficient algorithm for MINMTC in trees by using complementary slackness conditions: can it be done for MAXIMTF?

Actually, one can prove that the tree inequalities for MAXIMTF are a special case of a more general class of valid inequalities, that we call *inner odd set inequalities*: they have been used very recently (the paper [17] has appeared just after the submission of the present paper) to prove that Undirected MAXIMTF was polynomial-time solvable via the ellipsoid method. They are derived from the fact that evenness considerations are well-known to be of great importance in integer flow problems with several sources and sinks (see [12] for example). They can be defined as follows:

Definition 1. Let $G = (V, E)$ be an undirected graph and let T be the set of terminal vertices. For each $X \subseteq V \setminus T$, let $(X, V \setminus X)$ be the set of edges lying between X and $V \setminus X$ and let $c(X, V \setminus X)$ be the total capacity of the edges in $(X, V \setminus X)$. Then, F , the total number of flow units routed between the terminals in T , satisfies the following valid inequality (called inner odd set inequality)

$$F \leq \left\lfloor \frac{c(X, V \setminus X)}{2} \right\rfloor .$$

Roughly speaking, the validity of these inequalities comes from the fact that each flow unit in G is routed through no or an even number of edges in $(X, V \setminus X)$, since X contains no terminal (see Fig. 4). Hence, each flow unit routed through an edge in $(X, V \setminus X)$ is counted at least twice, and the total amount of flow in $(X, V \setminus X)$ is thus equal to (at least) $2F$. This amount being at most $c(X, V \setminus X)$, we can divide both sides of the inequality by 2 and use the integrality of F to obtain the desired result.

Now we can prove Theorem 6 by using Definition 1 and the following fact.

Theorem 7. *The tree inequalities are a special case of the inner odd set inequalities.*

Proof. We use the notations of Theorem 6. Given an undirected tree U , for each i , let (u_i, v_i) be an edge of p_i with capacity c_i (u_i lying in the path from t_i to v_i),

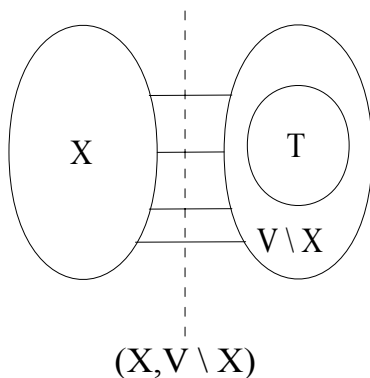


Fig. 4. An inner odd set configuration ($c(X, V \setminus X)$ is odd)

and let p'_i be the path from t_i to u_i (see Fig. 3). Then, the tree inequality on U is obtained by taking the inner odd set inequality defined on the set $X = V \setminus \bigcup_i p'_i$, since any flow unit has to be routed through at least one edge (and, in fact, through exactly two edges) in $(X, V \setminus X)$, i.e., through (u_i, v_i) and (u_j, v_j) for some i and j . \square

In trees, it is not difficult to see that all the inner odd set inequalities that matters are the ones corresponding to tree inequalities. Hence, from [17], these inequalities suffice to guarantee the existence of integer optimal solutions in undirected trees.

In fact, there exists an interesting relationship between the inner odd set inequalities and the inner eulerian assumption made in [12]. Given an edge-capacitated undirected graph $G = (V, E)$, the *degree* of a vertex $v \in V$, denoted by $d_V(v)$, is the sum of the capacities of the edges adjacent to v . Moreover, for $X \subseteq V$, $d_X(v)$ is the degree of vertex v in the subgraph of G induced by $X \cup \{v\}$. A graph is *inner eulerian* if every non terminal vertex has an even degree. Theorem 8 shows that the inner odd set inequalities are useless if the graph is inner eulerian.

Theorem 8. *Given a graph $G = (V, E)$ and a set of terminal vertices T , G is inner eulerian if and only if $\forall X \subseteq V \setminus T$, $c(X, V \setminus X)$ is even.*

Proof. It is easily seen that, by definition, G is inner eulerian if and only if $\forall X \subseteq V \setminus T$ with $|X| = 1$, $c(X, V \setminus X)$ is even. Now, assume that G is inner eulerian, and let $X \subset V$ contain no terminal. Then, we have

$$c(X, V \setminus X) = \sum_{v \in X} d_{V \setminus X}(v) = \sum_{v \in X} d_V(v) - \sum_{v \in X} d_X(v) .$$

Moreover, $\sum_{v \in X} d_V(v)$ is even since G is inner eulerian, and $\sum_{v \in X} d_X(v)$ is always even (because it is equal to two times the sum of the capacities of the

edges having both endpoints in X). This implies that $c(X, V \setminus X)$ is even, and Theorem 8 follows. \square

An interesting question would be to determine whether the inner odd set inequalities (together with the usual constraints) give a complete characterization of the polytope of Undirected MaxIMTF. Theorem 8 shows that a positive answer to this question would imply that the polytope of the continuous relaxation of the LP formulation of Undirected MaxIMTF is integral in inner eulerian undirected graphs (the existence of integer optimal solutions was already known [12]).

5 Conclusion and Open Problems

The parameter k_L introduced in this paper improves our knowledge of the boundary between tractable and intractable cases of MAXIMTF in directed graphs. Several results, positive and negative, about tractability and approximability of this problem, are provided. Moreover, we have given a family of valid inequalities for the undirected case, and have proved an interesting correspondence with valid inequalities already known. However, two important questions remain open: is there an $O(1)$ -approximation algorithm for the general directed case? Moreover, are there other tractable special cases (e.g., planar digraphs)?

References

1. Ahuja, A.K., Magnanti, T.L., Orlin, J.B.: Network Flows – Theory, Algorithms, and Applications. Prentice Hall, Englewood Cliffs, New Jersey (1993).
2. Andrews, M., Zhang, L.: Hardness of the undirected edge-disjoint paths problem. Proceedings STOC'05 (2005) 276–283.
3. Bertsimas, D., Teo, C.-P., Vohra, R.: Analysis of LP relaxations for multiway and multicut problems. Networks **34** (1999) 102–114.
4. Billionnet, A., Costa, M.-C.: Multiway cut and integer flow problems in trees. In: Liberti, L., Maffioli, F. (eds.): CTW04 Workshop on Graphs and Combinatorial Optimization. Electronic Notes in Discrete Mathematics **17** (2004) 105–109.
5. Călinescu, G., Karloff, H., Rabani, Y.: An improved approximation algorithm for Multiway Cut. Proceedings STOC'98 (1998) 48–52.
6. Chen, D.Z., Wu, X.: Efficient algorithms for k -terminal cuts on planar graphs. Algorithmica **38** (2004) 299–316.
7. Chopra, S., Rao, M.R.: On the multiway cut polyhedron. Networks **21** (1991) 51–89.
8. Costa, M.-C., Létocart, L., Roupin, F.: Minimal multicut and maximal integer multiflow: a survey. European Journal of Operational Research **162** (2005) 55–69. Elsevier.
9. Dahlhaus, E., Johnson, D.S., Papadimitriou, C.H., Seymour, P.D., Yannakakis, M.: The complexity of multiterminal cuts. SIAM Journal On Computing **23** (1994) 864–894.
10. Even, S., Itai, A., Shamir, A.: On the complexity of timetable and multicommodity flow problems. SIAM Journal on Computing **5** (1976) 691–703.

11. Ford, L.R., Fulkerson, D.R.: Maximal Flow Through a Network. *Canadian Journal of Mathematics* **8** (1956) 339–404.
12. Frank, A., Karzanov, A., Sebö, A.: On integer multiflow maximization, *SIAM J. Discrete Mathematics* **10** (1997) 158–170.
13. Garg, N., Vazirani, V.V., Yannakakis, M.: Multiway cuts in directed and node weighted graphs. In: Abiteboul, S., Shamir, E. (eds.): 21st International Colloquium on Automata, Languages and Programming. *Lecture Notes in Computer Science*, Vol. 820. Springer-Verlag (1994) 487–498.
14. Garg, N., Vazirani, V.V., Yannakakis, M.: Primal-dual approximation algorithms for integral flow and multicut in trees. *Algorithmica* **18** (1997) 3–20.
15. Guruswami, V., Khanna, S., Rajaraman, R., Shepherd, B., Yannakakis, M.: Near-optimal hardness results and approximation algorithms for edge-disjoint paths and related problems. *Proceedings STOC'99* (1999) 19–28.
16. Karger, D., Klein, P., Stein, C., Thorup, M., Young, N.: Rounding Algorithms for a Geometric Embedding of Minimum Multiway Cut. *Proceedings STOC'99* (1999) 668–678.
17. Keijsper, J.C.M., Pendavingh, R.A., Stougie, L.: A linear programming formulation of Mader's edge-disjoint paths problem. *Journal of Combinatorial Theory, Series B* **96** (2006) 159–163.
18. Naor, J., Zosin, L.: A 2-approximation algorithm for the directed multiway cut problem. *Proceedings FOCS'97* (1997) 548–553.
19. Schrijver, A.: *Combinatorial Optimization - Polyhedra and Efficiency*. Algorithms and Combinatorics **24**. Springer-Verlag (2003).
20. Yeh, W.-C.: A Simple Algorithm for the Planar Multiway Cut Problem. *Journal of Algorithms* **39** (2001) 68–77.

Routing with Early Ordering for Just-In-Time Manufacturing Systems

Mingzhou Jin, Kai Liu, and Burak Eksioglu

Department of Industrial Engineering, Mississippi State University,
P.O. Box 9542, Mississippi State, MS, 39762, USA
mj@ie.msstate.edu

Abstract. Parts required in Just-In-Time manufacturing systems are usually picked up from suppliers on a daily basis, and the routes are determined based on average demand. Because of high demand variance, static routes result in low truck utilization and occasional overflow. Dynamic routing with limited early ordering can significantly reduce transportation costs. An integrated mixed integer programming model is presented to capture transportation cost, early ordering inventory cost and stop cost with the concept of rolling horizon. A four-stage heuristic algorithm is developed to solve a real-life problem. The stages of the algorithms are: determining the number of trucks required, grouping, early ordering, and routing. Significant cost savings is estimated based on real data.

1 Introduction

The Just-In-Time (JIT) philosophy originated from the work of Taiichi Ohno at Toyota Motor Company and made its way to the US about 20 years ago [1]. The JIT philosophy is now adopted by most automakers all over the world. Based on the JIT philosophy, inventory is considered a big cost contributor so that the goal is to reduce inventory levels to “zero” [2]. Therefore, parts are ordered and transported only when they are needed in the production. Based on a recent project with one of the major automakers in the US that implements a JIT system, we found out that the inbound logistics decision making process has the following procedures:

1. The production plan is determined based on dealer orders or forecasted demand, and it is derived by manufacturing needs such as line balancing.
2. The parts are ordered based on daily production needs, and the logistics group has little control on how many to order and when to order.
3. Milk-runs (routes) are determined based on average demand. Following the milk-runs trucks visit the suppliers to pick up the required parts and then come back to the manufacturing plant. Each truck has one run everyday. A supplier may be visited by one or more trucks. This is different from the typical capacitated vehicle routing problem (CVRP), in which each supplier is visited by exactly one truck. The problem is similar to the split delivery vehicle routing problem (SDVRP) [3, 4].

Because inbound logistics costs are not considered during production planning, high volatility of part consumption in the assembly line leads to frequent and small

batches. The practice of assigning the suppliers to the trucks (milk-runs) based on average daily demand and keeping the same routes every day makes the problem worse. Furthermore, routes are typically determined manually, which also hurts the efficiency. Truck utilizations can be as low as 30%, while sometimes overflow happens when additional trucks are required.

In this paper we address the above mentioned problem of the mismatch between production and logistics. To eliminate this mismatch, “dynamic routing” and “early ordering” are proposed as solutions. Dynamic routing means that the routes are determined daily based on production needs. Though integrated production planning and route scheduling is implemented in some other industries [5], in the automotive industry it is difficult to fully incorporate logistics needs into production planning. The industry has implemented a manufacturing driven JIT system for such a long time that any major change requires approval from high-level management, which is usually difficult to achieve. The proposed early ordering policy will not affect the production planning process. In the automotive industry, production plans are made several days before actual production starts. Thus, the required parts are known several days before they are actually consumed in the assembly line. Early ordering policy simply allows the parts to be shipped one or two days early to save transportation cost. However, late ordering is not allowed in order not to disturb the manufacturing process. The parts that are ordered early will be stored in the inbound warehouse for one or two days and possibly increase inventory holding costs.

In Section 2, a mixed integer programming (MIP) model is proposed for the routing problem with early ordering in a JIT environment. A heuristic algorithm to solve the model is presented in Section 3. Section 4 concludes the paper and includes some cost saving estimates based on a real case.

2 A Mixed Integer Programming Model for Daily Routing with Early Ordering

In an assembly system such as the auto assembly plant described above, there are two main costs: transportation and early ordering inventory costs. The transportation cost is composed of a fixed cost for each truck, a variable cost for each mile, and a fixed cost for each stop. Among them, the fixed cost for each truck dominates the others. In the literature, most routing studies do not consider the stop issue. The automotive company that we have studied has to pay a fixed amount to the trucking companies for each stop on the routes because a stop means additional handling time and effort. The number of stops is also a constraint because a truck can not finish a route in one day if it has to make too many stops. In the standard CVRP models, since each supplier can be visited exactly once, the number of total stops is fixed. Therefore, there is no need to consider stop costs in the CVRP models. Though the routing decision influences the number of stops in an SDVRP model, stop costs and constraints are usually not addressed in the SDVRP literature [6-8].

Since early ordering is allowed, additional inventory holding costs are considered in the proposed model. Typically, a major component of the inventory holding cost is the cost of capital invested [1]. However, automakers and their suppliers have long-term relationships, and the payments are made periodically (e.g. weekly or biweekly).

Therefore, the inventory holding costs are mainly driven by the space occupied rather than the capital invested. The demand for one part could be several hundred pieces or more every day, and they are held in containers during shipping and handling. The number of parts in a container is called a unit load, which may have several dozen or up to hundred pieces of the same part. Therefore, the demand and the amount delivered are measured in unit loads rather than pieces. The notation and the model for the routing problem with early ordering are given below.

Parameters:

- K : the number of available trucks ($k=1,2,\dots,K$);
- T : the number of days in the planning horizon ($t=1,2,\dots,T$);
- P : the set of parts ($p \in P$);
- N : the number of suppliers ($i,j=0,1,2,\dots,N$; 0 is used for the origin);
- C : truck capacity;
- u : the inventory holding cost per unit space for early ordered parts;
- q : the fixed cost per truck per day;
- λ : the variable transportation cost per truck per mile;
- w : the cost for one stop;
- $d_{t,p}$: the demand (in unit loads) for part p on day t (with lead time);
- $c_{i,j}$: the distance from supplier i to supplier j ;
- $r_{p,i}$: indicates whether or not part p is provided by supplier i (0: no; 1: yes; note that each part is provided only by one supplier; $\sum_{i=1}^N r_{p,i} = 1, p \in P$);
- v_p : space required by one unit load of part p ;
- S : the maximum number of stops allowed for each truck.

Decision variables:

- $o_{t,k,p}$: the unit loads of part p shipped by truck k on day t ;
- $x_{t,k,i,j}$: equals 1 if truck k visits supplier j right after supplier i on day t ; 0 otherwise;
- $l_{t,k,i}$: the remaining capacity of truck k after visiting supplier i on day t ($l_{0,k,t} = C$);
- $s_{t,k,i}$: equals 1 if truck k visits supplier i on day t ; 0 otherwise;
- $I_{t,p}$: the inventory (in unit loads) of p ordered early on day t .

The MIP model for daily routing with early ordering:

$$\min u \sum_{t=1}^T \sum_{p \in P} v_p I_{t,p} + q \sum_{t=1}^T \sum_{k=1}^K \sum_{j=1}^N x_{t,k,0,j} + \lambda \sum_{t=1}^T \sum_{k=1}^K \sum_{i,j=0}^N c_{i,j} x_{t,k,i,j} + w \sum_{t=1}^T \sum_{k=1}^K \sum_{j=1}^N s_{t,k,j} \quad (1.1)$$

$$s.t. \quad I_{t-1,p} - I_{t,p} + \sum_{k=1}^K o_{t,k,p} = d_{t,p} \quad t = 1, \dots, T; p \in P; \quad (1.2)$$

$$-C \cdot s_{t,k,i} + \sum_{p \in P} v_p r_{p,i} o_{t,k,p} \leq 0 \quad t = 1, \dots, T; k = 1, \dots, K; i = 1, \dots, N; \quad (1.3)$$

$$\sum_{i=0, \dots, N, i \neq j} x_{t,k,i,j} - s_{t,k,j} = 0 \quad t = 1, \dots, T; k = 1, \dots, K; j = 1, \dots, N; \quad (1.4)$$

$$\sum_{i=0, \dots, N, i \neq j} x_{t,k,i,j} - \sum_{i=0, \dots, N, i \neq j} x_{t,k,j,i} = 0 \quad t = 1, \dots, T; k = 1, \dots, K; j = 1, \dots, N; \quad (1.5)$$

$$Cx_{t,k,i,j} + \sum_{p \in P} v_p r_{p,i} o_{t,k,p} - l_{t,k,i} + l_{t,k,j} \leq C \quad t = 1, \dots, T; k = 1, \dots, K; \quad (1.6)$$

$$i, j = 0, \dots, N; j \neq i;$$

$$\sum_{j=1}^N s_{t,k,j} \leq S \quad t = 1, \dots, T; k = 1, \dots, K; \quad (1.7)$$

$$l_{t,k,i}, I_{t,p}, o_{t,k,p} \geq 0; s_{t,k,i}, x_{t,k,i,j} : \text{binary}; o_{t,k,p} : \text{integers}.$$

The objective function (1.1) minimizes the sum of the inventory holding costs and the transportation costs. The first constraint set (1.2) represents the inventory evolvment over days and ensures that the production needs of all parts are satisfied. The second constraint set (1.3) ensures that only those trucks that visit supplier i pick up parts from supplier i . The third constraint set (1.4) help obtain the numbers of truck stops. The fourth constraint set (1.5) is used to keep the flow at each supplier balanced (i.e. the number of trucks arriving at a supplier equals the number of trucks leaving that supplier). The fifth constraint set (1.6) makes sure that the amount picked up by a truck does not exceed its capacity and eliminates sub-tours. The last set of constraints (1.7) is used to make sure that the number of stops by a truck does not exceed the maximum number allowed.

The proposed model combines transportation and inventory decisions. It is a variant of the SDVRP with additional constraints to address the issues of truck stops and early ordering. The standard SDVRP takes the total required space at each demand point into consideration, but this model addresses the problem at the part level. Since thousands of parts are required and hundreds of suppliers need to be visited every day, the model of the whole inbound logistics for this automaker is very large. The SDVRP itself is a well-known *NP-complete* problem [3]. Therefore, good solutions rather than optimal solutions are expected in practice for a large-scale SDVRP. The early ordering policy makes the computational burden even heavier by adding one more dimension of time into the problem. In reality, the suppliers are usually grouped by regions, and several transportation service providers are in charge of one region. Thus, the logistics problem of each region can be solved separately. However, dozens of suppliers and hundreds of parts are usually involved in a region. In a problem with 15 suppliers, 150 parts, and 7 available trucks there will be more than 3000 binary variables when only one-day early ordering is allowed. When CPLEX [9], a commercial optimization software, is used to solve the model, 800MB of memory is used up after one day of calculations while the gap between the lower and upper bounds is still more than 90%. Therefore, fast heuristics are necessary for daily decision making.

3 A Four-Stage Heuristic Algorithm

A number of constructive heuristics can be found in the literature for the CVRP or the SDVRP. Most of them are two-stage algorithms including a grouping stage and a routing stage. Typically, a bin packing problem with restrictions is used in the grouping stage. Belenguer et al. [4] develop an algorithm based on a lower bound to

solve the SDVRP (this needs some more explanation). Archetti et al. [5] use dynamic programming to solve SDVRP instances where vehicles have small capacities. For the model given in Section 2, we develop a four-stage heuristic algorithm. The basic scheme is illustrated in Fig. 1.

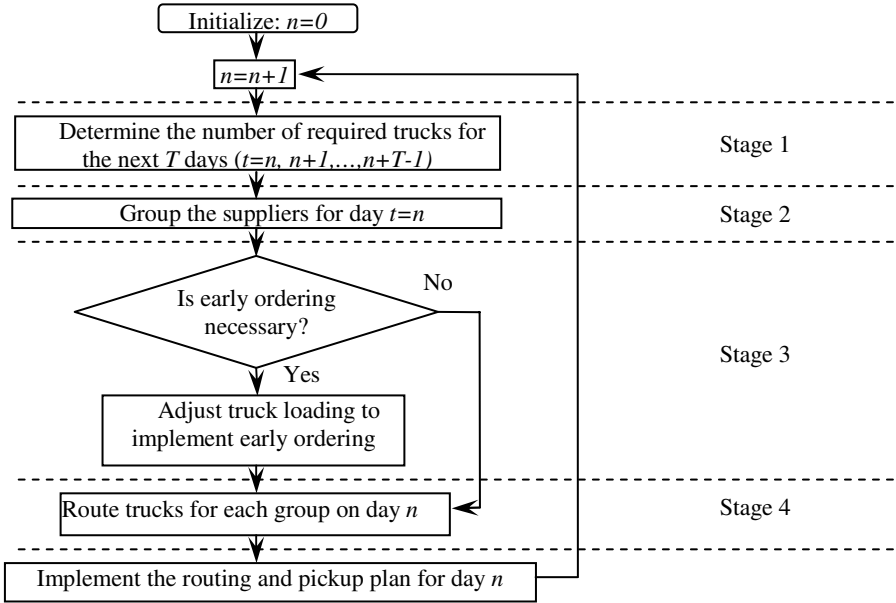


Fig. 1. The basic scheme of the four-stage heuristic

In the first stage, the number of required trucks is calculated for days $t=n, n+1, \dots, n+T-1$. Initially, early ordering is not considered and the number of required trucks is found by the following equation:

$$K_t = \frac{\theta_t (\text{Total required volume of day } t)}{C (\text{truck capacity}) \times 0.9 (\text{Allowance})}, \tag{2}$$

where $\theta_t = \sum_{p \in P} v_p d_{t,p}$ and the allowance of 0.9 is used to account for space that may

be wasted because of unit loads. The number of required trucks on day t without early ordering is $\lceil K_t \rceil$. If only one-day early ordering is allowed, early ordering is implemented if the following two conditions are satisfied:

$$\lceil K_t + K_{t+1} \rceil < \lceil K_t \rceil + \lceil K_{t+1} \rceil; \tag{3.1}$$

$$\text{and } K_{t+1} - \lfloor K_{t+1} \rfloor \leq UB. \tag{3.2}$$

Since the fixed cost for each truck dominates other costs, early ordering is only implemented when a truck can be saved on the next day (condition (3.1)). For example, if $K_t=2.3$ and $K_{t+1}=2.2$ then three trucks are required for both days without

early ordering. If 0.2 truck space of parts can be moved from day $t+1$ to day t (i.e. $K_t=2.5$ and $K_{t+1}=2$), one truck can be saved on day $t+1$. Also, note that early ordering increases inventory holding costs. Therefore, it is implemented only when the space moved from day $t+1$ to day t is not more than UB (e.g. 30%) of the truck capacity (condition (3.2)).

In the second stage, grouping is done to determine which suppliers should be visited by each truck. There are many grouping heuristics in the literature. The basic idea is to group the nearby suppliers together without violating the truck capacity constraints. We propose the following optimization model to solve the grouping problem for each day t .

$$\min w \sum_{i=1}^N \sum_{k=1}^K s_{t,k,i} + \tau \sum_{i=1}^N \sum_{k=1}^K (J_{i,k}^{x+} + J_{i,k}^{x-} + J_{i,k}^{y+} + J_{i,k}^{y-}) \tag{4.1}$$

$$s.t. \sum_{p \in P} v_p o_{t,k,p} \leq C \quad k = 1, 2, \dots, K; \tag{4.2}$$

$$\sum_{k=1}^K o_{t,k,p} = d_{t,p} \quad p \in P; \tag{4.3}$$

$$\sum_{p \in P} v_p r_{p,i} o_{t,k,p} \leq C s_{t,k,i} \quad k = 1, 2, \dots, K; i = 1, 2, \dots, N; \tag{4.4}$$

$$\sum_{i=1}^N s_{t,k,i} \leq S \quad k = 1, 2, \dots, K; \tag{4.5}$$

$$q_k^x - b_i^x \leq M(1 - s_{t,k,i}) + J_{i,k}^{x+} \quad i = 1, 2, \dots, N; k = 1, 2, \dots, K; \tag{4.6}$$

$$b_i^x - q_k^x \leq M(1 - s_{t,k,i}) + J_{i,k}^{x-} \quad i = 1, 2, \dots, N; k = 1, 2, \dots, K; \tag{4.7}$$

$$q_k^y - b_i^y \leq M(1 - s_{t,k,i}) + J_{i,k}^{y+} \quad i = 1, 2, \dots, N; k = 1, 2, \dots, K; \tag{4.8}$$

$$b_i^y - q_k^y \leq M(1 - s_{t,k,i}) + J_{i,k}^{y-} \quad i = 1, 2, \dots, N; k = 1, 2, \dots, K; \tag{4.9}$$

$$s_{t,k,i} : \text{binary}; o_{t,k,p} \geq 0, \text{ integer}; q_k^x, q_k^y, J_{i,k}^{x+}, J_{i,k}^{x-}, J_{i,k}^{y+}, J_{i,k}^{y-} \geq 0.$$

where

- b_i^x, b_i^y : the (x, y) coordinators of supplier i ;
- M : a large number to facilitate modeling;
- τ : cost per unit distance;
- q_k^x, q_k^y : the latitude and longitude of the virtual center for truck k ;
- $J_{i,k}^{x+}, J_{i,k}^{x-}, J_{i,k}^{y+}, J_{i,k}^{y-}$: the rectangular distance between supplier i and virtual center of truck k .

If truck k serves a set of suppliers (I_k) in the grouping model, its virtual center is defined as (q_k^x, q_k^y) that minimizes $\sum_{i \in I_k} (|b_i^x - q_k^x| + |b_i^y - q_k^y|)$. The objective function (4.1) minimizes the total costs including the stop cost and the distance cost

approximated by the sum of rectangular distances between the virtual center of the trucks and the suppliers served by those trucks. The first constraint set (4.2) forces the load of each truck to be less than or equal to the capacity. Constraint set (4.3) has demand satisfied for all parts requested by the assembly line. Constraint set (4.4) ensures that truck k stops at supplier i if parts are to be picked by truck k from that supplier. Constraint set (4.5) makes sure that the number of stops by a truck does not exceed the maximum number allowed. The last four constraint sets (4.6-4.9) are used to obtain the rectangular distances from the suppliers to the virtual centers of the trucks. Though the model looks cumbersome, the number of variables and constraints are significantly less than those of the original model (1.1-1.7). A problem with 15 suppliers, 150 parts and 7 available trucks has about 105 binary variables and 750 constraints. CPLEX can yield a solution to such a relatively small problem in 5 minutes with less than 1% gap on average.

In the third stage, how to implement the early ordering policy is determined. Let $L_{k,t}$ be the remaining capacity of truck k on day t . The following algorithm given in Fig. 2 is developed for implementing one-day early ordering. Assume e_i is the total space occupied by the parts ordered early on day t ($e_i = 0.9C(K_{t+1} - \lfloor K_{t+1} \rfloor)$).

Step 0. Initialize k ($k = 1$)

Step 1. Move all parts provided by one supplier i , served by truck k and satisfying both of the following conditions:

- they are needed on both day t and $t+1$:

$$\sum_{p \in P} r_{p,i} o_{t,k,p} \geq 0 \quad \text{and} \quad \sum_{p \in P} r_{p,i} d_{t+1,p} \geq 0 ;$$

- the total day $t+1$ volume from the supplier i does not exceed the remaining capacity: $\sum_{p \in P} r_{p,i} d_{t+1,p} \leq C - L_k$;

If a movement happens,

- update the utilized capacity L_k of truck k ;
- update the total early ordered volume.

Step 3. If the total early ordered volume reaches e_i , go to end.

Step 4. If $k < K$: $k=k+1$ and go to step 1.

Step 5. Move the parts satisfying the following conditions for truck $k=1, \dots, K$ until the total early ordered volume reaches e_i :

- It is needed on both day t and $t+1$: $o_{t,k,p} \geq 0$ and $d_{t+1,p} \geq 0$
- Day $t+1$ volume of the same part does not exceed the remaining capacity of the truck: $d_{t+1,p} \leq C - L_k$

Step 6. If the total early ordered volume is still smaller than e_i , arbitrarily move the parts needed on both days until the early ordered volume reaches e_i .

Fig. 2. The Algorithm to Determine the Early Ordering Policy

The algorithm that determines the early ordering policy first tries to reduce the number of stops on day $t+1$ without increasing the number of stops on day t . The second priority is to reduce the number of handlings of the same parts.

The fourth stage deals with the routing problem for each truck which is a standard Travel Sales Problem (TSP). Though the TSP is an *NP-hard* problem, its optimal solution can be obtained in seconds by CPLEX in this case because each truck usually has at most five stops.

The concept of rolling horizon is used for the overall algorithm. If only one-day early ordering is allowed ($T=2$), the first step is implemented for two days ($t=n$ and $n+1$). Grouping and routing models are only solved for the first day ($t=n$). On the next day, the second day's demand information will be updated, and the early ordered parts will be deducted from it. The four-stage algorithm will be implemented after updating $n=n+1$ with the new information of the third day's demand.

4 Implementation and Conclusion

The proposed four-stage algorithm is implemented on a real inbound logistics problem faced by a major automotive company in the US. The region under study has 15 suppliers and 158 parts, and usually 4 to 7 trucks are required every day. Only one-day early ordering is allowed because of information availability and inventory concerns. One month's worth of real data is used, and the result is obtained in 10 minutes. The average truck utilization is improved from 40% to 80% while its variability over trucks also becomes much smaller. The total cost savings is about 20% including 24% savings on the number of used trucks, 17% savings on the total number of stops, and 15% savings on the total traveled distance. Early ordering does not happen frequently in that about 2% of the parts (in space) are ordered early. Of the total 20% savings, about 4% is contributed by the early ordering policy.

This paper presents a mixed integer programming model and a heuristic algorithm to improve the inbound logistics for Just-In-Time manufacturing systems. A dynamic routing policy and an early ordering policy are proposed to reduce the total cost. The proposed model and the algorithm are tested using real-life data obtained from an auto manufacturer. The recommended policies and the proposed heuristic algorithm works well in the sense that the computational speed is high while the quality of the solution is much better than that of the solution currently used by the auto manufacturer.

In this paper, only a constructive heuristic is discussed without any improvement phase. A future research extension is to develop more sophisticated improvement heuristics to obtain better solutions to the problem.

References

1. Askin, R., Golderg, J.: Design and Analysis of Lean Manufacturing. John Wiley & Sons, New York (2002).
2. Monden, Y.: Toyota production system - 3rd edition: An Integrated Approach to Just-In-Time. Institute of Industrial Engineering, Norcross, GA (1993).
3. Dror, M., Trudeau, P.: Savings by Split Delivery Routing. Transportation Science 23 (1989) 141-145.

4. Dror, M., Laporte, G., Trudeau P.: Vehicle Routing with Split Delivery. *Discrete Applied Mathematics* 50 (1994) 239-254.
5. Bredström, D., Rönnqvist, M.: Integrated Production Planning and Route Scheduling in Pulp Mill Industry. *Proceedings of the 35th Hawaii International Conference on System Sciences* (2002).
6. Belenguer, J.M., Martinez, M.C., Mota, E.: Lower Bound for the Split Delivery Vehicle Routing Problem. *Operations Research* 48 (2000) 801-810.
7. Archetti, C., Mansini, R., Speranza, M.G.: The Split Delivery Vehicle Routing Problem with Small Capacity. Technical Report, Department of Quantitative Methods, University of Brescia (2001).
8. Lee, C.G., Epelman, M., White, C.C.: A Shortest Path Approach to the Multiple-vehicle Routing with Split Picks-up. To appear in *Transportation Research - Part B* (2005).
9. ILOG, ILOG CPLEX 9.0 User's Manual, 2003.

A Variant of the Constant Step Rule for Approximate Subgradient Methods over Nonlinear Networks*

Eugenio Mijangos

University of the Basque Country, Department of Applied Mathematics and Statistics and Operations Research, P.O. Box 644, 48080 Bilbao, Spain
<http://www.ehu.es/~mepmifee>

Abstract. The efficiency of the network flow techniques can be exploited in the solution of nonlinearly constrained network flow problems (NCNFP) by means of approximate subgradient methods (ASM). We propose to solve the dual problem by an ASM that uses a variant of the well-known constant step rule of Shor. In this work the kind of convergence of this method is analyzed and its efficiency is compared with that of other approximate subgradient methods over NCNFP.

1 Introduction

Many nonlinear network flow problems have nonlinear side constraints, these are named nonlinearly constrained network flow problems (**NCNFP**) and can be expressed as

$$\underset{x}{\text{minimize}} \quad f(x) \tag{1}$$

$$\text{subject to} \quad x \in \mathcal{F} \tag{2}$$

$$c(x) \leq 0, \tag{3}$$

where:

- The set \mathcal{F} is

$$\mathcal{F} = \{x \in \mathbb{R}^n \mid Ax = b, 0 \leq x \leq \bar{x}\},$$

where A is a node-arc incidence $m \times n$ -matrix, b is the production/demand m -vector, x are the flows on the arcs of the network represented by A , and \bar{x} are the capacity bounds imposed on the flows of each arc.

- The side constraints (3) are defined by $c : \mathbb{R}^n \rightarrow \mathbb{R}^r$, such that $c = [c_1, \dots, c_r]^t$, where $c_i(x)$ is nonlinear and twice continuously differentiable on the feasible set \mathcal{F} for all $i = 1, \dots, r$.
- $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is nonlinear and twice continuously differentiable on \mathcal{F} .

* The research was partially supported by grant MCYT DPI 2005-09117-C02-01.

In recent works **NCNFP** has been solved using partial augmented Lagrangian methods with quadratic penalty function [7] and with exponential penalty function [8], and using approximate subgradient methods [8, 9].

In this work we focus on the primal problem **NCNFP** and its dual problem

$$\text{maximize } q(\mu) = \min_{x \in \mathcal{F}} l(x, \mu) = \min_{x \in \mathcal{F}} \{f(x) + \mu^t c(x)\} \quad (4)$$

$$\text{subject to: } \mu \in \mathcal{M}, \quad (5)$$

where $\mathcal{M} = \{\mu \mid \mu \geq 0, q(\mu) > -\infty\}$. We assume throughout this paper that the constraint set \mathcal{M} is closed and convex, q is continuous on \mathcal{M} , and for every $\mu \in \mathcal{M}$ some vector $x(\mu)$ that minimizes $l(x, \mu)$ over $x \in \mathcal{F}$ can be calculated, yielding a subgradient $c(x(\mu))$ of q at μ . We propose to solve **NCNFP** by using primal-dual methods, see [1].

The minimization of the Lagrangian function $l(x, \mu)$ over \mathcal{F} can be performed by means of efficient techniques specialized for networks, see [14].

Since $q(\mu)$ is approximately computed, we consider *approximate subgradient methods* [9] in the solution of this problem. The basic difference between these methods and the classical subgradient methods is that they replace the subgradients with inexact subgradients.

Different ways of computing the stepsize in the approximate subgradient methods have been considered. The diminishing stepsize rule (DSR) suggested by Correa and Lemaréchal in [3] for exact subgradients. A dynamically chosen stepsize rule that uses an estimation of the optimal value of the dual function by means of an adjustment procedure (DSRAP) similar to that suggested by Nedić and Bertsekas in [11] for incremental subgradient methods. A dynamically chosen stepsize whose estimate of the optimal value of the dual function is based on the relaxation level-control algorithm (DSRLC) designed by Brännlund in [2] and analyzed by Goffin and Kiwiel in [5]. A variant of the constant step rule (VCSR) of Shor [13].

The convergence of the first three methods was studied in the cited papers for the case of exact subgradients. The convergence of the corresponding approximate (inexact) subgradient methods is analyzed in [9], see also [6].

In this work the convergence of VCSR is analyzed and its efficiency compared with that of DSR, DSRAP, and DSRLC over **NCNFP** problems.

The remainder of this paper is structured as follows. Section 2 presents the variant of the constant step rule together with the other ways of computing the stepsize in the approximate subgradient methods; Section 3 describes the solution to the nonlinearly constrained network flow problem; and Section 4 puts forward experimental results. Finally, Section 5 concludes the paper.

2 Approximate Subgradient Methods

When, as happens in this work, for a given $\mu \in \mathcal{M}$, the dual function value $q(\mu)$ is calculated by minimizing approximately $l(x, \mu)$ over $x \in \mathcal{F}$ (see (4)), the subgradient obtained (as well as the value of $q(\mu)$) will involve an error.

In order to analyze such methods, it is useful to introduce a notion of approximate subgradient [13, 1]. In particular, given a scalar $\varepsilon \geq 0$ and a vector $\bar{\mu}$ with $q(\bar{\mu}) > -\infty$, we say that c is an ε -subgradient at $\bar{\mu}$ if

$$q(\mu) \leq q(\bar{\mu}) + \varepsilon + c^t(\mu - \bar{\mu}), \quad \forall \mu \in \mathbb{R}^r. \tag{6}$$

The set of all ε -subgradients at $\bar{\mu}$ is called the ε -subdifferential at $\bar{\mu}$ and is denoted by $\partial_\varepsilon q(\bar{\mu})$.

An approximate subgradient method is defined by

$$\mu^{k+1} = [\mu^k + s_k c^k]^+, \tag{7}$$

where c^k is an ε_k -subgradient at μ^k and s_k a positive stepsize.

In our context, we minimize approximately $l(x, \mu^k)$ over $x \in \mathcal{F}$, thereby obtaining a vector $x^k \in \mathcal{F}$ with

$$l(x^k, \mu^k) \leq \inf_{x \in \mathcal{F}} l(x, \mu^k) + \varepsilon_k. \tag{8}$$

As is shown in [1, 9], the corresponding constraint vector, $c(\mu^k)$, is an ε_k -subgradient at μ^k . If we denote $q_{\varepsilon_k}(\mu^k) = l(x^k, \mu^k)$, by definition of $q(\mu^k)$ and using (8) we have

$$q(\mu^k) \leq q_{\varepsilon_k}(\mu^k) \leq q(\mu^k) + \varepsilon_k \quad \forall k. \tag{9}$$

2.1 Stepsize Rules

Throughout this paper, we use the notation

$$q^* = \sup_{\mu \in \mathcal{M}} q(\mu), \quad \mathcal{M}^* = \{\mu \in \mathcal{M} \mid q(\mu) = q^*\}, \quad \text{dist}(\mu, \mathcal{M}^*) = \inf_{\mu^* \in \mathcal{M}^*} \|\mu - \mu^*\|,$$

where $\|\cdot\|$ denotes the standard Euclidean norm.

Assumption 1: There exists scalar $C > 0$ such that for $\mu^k \in \mathcal{M}$, $\varepsilon_k \geq 0$ and $c^k \in \partial_{\varepsilon_k} q(\mu^k)$, we have $\|c^k\| \leq C$, for $k = 0, 1, \dots$

In this paper, four kinds of stepsize rules have been considered.

Variant of the Constant Step Rule (VCSR). As is well known the classical scaling of Shor (see [13])

$$s_k = \frac{s}{\|c^k\|}$$

gives rise to a s -constant-step algorithm.

Note that constant stepsizes (i.e., $s_k = s$ for all k) are unsuitable because the function q may be nondifferentiable at the optimal point and then $\{c^k\}$ does not necessarily tend to zero, even if $\{\mu^k\}$ converges to the optimal point, see [13].

On the other hand, in our case c^k is an approximate subgradient, hence it can exist a k such that $c^k \in \partial_{\varepsilon_k} q(\mu^k)$ with $\|c^k\| = 0$, but ε_k not being sufficiently small. In order to overcome this trouble we have considered the following variant:

$$s_k = \frac{s}{\delta + \|c^k\|}, \tag{10}$$

where s and δ are positive constants. The following proposition shows its kind of convergence.

Proposition 1. *Let Assumption 1 hold. Let the optimal set \mathcal{M}^* be nonempty. Suppose that a sequence $\{\mu^k\}$ is calculated by the ε -subgradient method given by (7), with the stepsize (10), where $\sum_{k=1}^\infty \varepsilon_k < \infty$. Then*

$$q^* - \limsup_{k \rightarrow \infty} q_{\varepsilon_k}(\mu^k) < \frac{\delta}{2}(\delta + C). \tag{11}$$

Proof. Let μ^* be a point that maximizes q ; i.e., $\mu^* \in \mathcal{M}^*$ such that $q^* = q(\mu^*)$. According to (7) we have

$$\begin{aligned} \|\mu^{k+1} - \mu^*\|^2 &\leq \|\mu^k + s_k c^k - \mu^*\|^2 \\ &= \|\mu^k - \mu^*\|^2 + 2s_k (c^k)^t (\mu^k - \mu^*) + s_k^2 \|c^k\|^2 \\ &\leq \|\mu^k - \mu^*\|^2 - 2s_k (q(\mu^*) - q_{\varepsilon_k}(\mu^k) - \varepsilon_k) + s_k^2 \|c^k\|^2, \end{aligned}$$

where the last inequality follows from the definition of ε -subgradient, see (6), which gives

$$q(\mu^*) \leq q_{\varepsilon_k}(\mu^k) + \varepsilon_k + (c^k)^t (\mu^* - \mu^k).$$

Applying the inequality above recursively, we have

$$\|\mu^{k+1} - \mu^*\|^2 \leq \|\mu^1 - \mu^*\|^2 - 2 \sum_{i=1}^k s_i [q(\mu^*) - q_{\varepsilon_i}(\mu^i)] + \sum_{i=1}^k (s_i^2 \|c^i\|^2 + 2\varepsilon_i),$$

which implies $2 \sum_{i=1}^k s_i [q(\mu^*) - q_{\varepsilon_i}(\mu^i)] \leq \|\mu^1 - \mu^*\|^2 + \sum_{i=1}^k (s_i^2 \|c^i\|^2 + 2\varepsilon_i)$. Combining this with

$$\sum_{i=1}^k s_i [q(\mu^*) - q_{\varepsilon_i}(\mu^i)] \geq \left(\sum_{i=1}^k s_i \right) [q(\mu^*) - \widehat{q}_{\varepsilon_k}],$$

where $\widehat{q}_{\varepsilon_k} = \max_{0 \leq i \leq k} q_{\varepsilon_i}(\mu^i)$, we have the inequality

$$2 \left(\sum_{i=1}^k s_i \right) [q(\mu^*) - \widehat{q}_{\varepsilon_k}] \leq \|\mu^1 - \mu^*\|^2 + \sum_{i=1}^k (s_i^2 \|c^i\|^2 + 2\varepsilon_i),$$

which is equivalent to

$$q(\mu^*) - \widehat{q}_{\varepsilon_k} \leq \frac{\|\mu^1 - \mu^*\|^2 + \sum_{i=1}^k (s_i^2 \|c^i\|^2 + 2\varepsilon_i)}{2 \sum_{i=1}^k s_i}.$$

Since μ^* is any maximizer of q , we can state that

$$q^* - \widehat{q}_{\varepsilon_k} \leq \frac{\text{dist}(\mu^1, \mathcal{M}^*)^2 + \sum_{i=1}^k s_i^2 \|c^i\|^2 + 2 \sum_{i=1}^k \varepsilon_i}{2 \sum_{i=1}^k s_i}. \tag{12}$$

On the other hand, as the stepsize is $s_i = s/(\delta + \|c^i\|)$, by Assumption 1 we have

$$\sum_{i=1}^k s_i > \frac{ks}{\delta + C} \tag{13}$$

and so $\sum_{i=1}^{\infty} s_i = \infty$. In addition, as $s_i^2 \|c^i\|^2 < s^2$, it holds $\sum_{i=1}^k s_i^2 \|c^i\|^2 < ks^2$. Therefore

$$\frac{\sum_{i=1}^k s_i^2 \|c^i\|^2}{\sum_{i=1}^k s_i} < \frac{ks^2}{\frac{ks}{\delta+C}} = s(\delta + C). \tag{14}$$

Taking into account the assumption $\sum_{k=1}^{\infty} \varepsilon_k < \infty$ together with (12)-(14), for $k \rightarrow \infty$ we obtain

$$q^* - \lim_{k \rightarrow \infty} \widehat{q}_{\varepsilon_k} \leq \frac{s}{2}(\delta + C), \tag{15}$$

as required. □

Note that for very small values of δ the stepsize (10) is similar to Shor’s classical scaling; in contrast, for big values of δ (with regard to $\sup\{\|c^k\|\}$) the stepsize (10) looks like a constant stepsize. In this work by default $\delta = 10^{-12}$, with $s = 100$.

Diminishing Stepsize Rule (DSR). The convergence of the exact subgradient method using a diminishing stepsize was shown by Correa and Lemaréchal, see [3].

In the case where c^k is an ε_k -subgradient we have the following proposition, which was proved in [9].

Proposition 2. *Let the optimal set \mathcal{M}^* be nonempty. Also, assume that the sequences $\{s_k\}$ and $\{\varepsilon_k\}$ are such that*

$$s_k > 0, \quad \sum_{k=0}^{\infty} s_k = \infty, \quad \sum_{k=0}^{\infty} s_k^2 < \infty, \quad \sum_{k=0}^{\infty} s_k \varepsilon_k < \infty. \tag{16}$$

Then, the sequence $\{\mu^k\}$, generated by the ε -subgradient method, where $c^k \in \partial_{\varepsilon_k} q(\mu^k)$ (with $\{\|c^k\|\}$ bounded), converges to some optimal solution.

An example of such a stepsize is

$$s^k = 1/\widehat{k},$$

for $\widehat{k} = \lfloor k/m \rfloor + 1$. We use by default $m = 5$.

An interesting alternative for the ordinary subgradient method is the *dynamic stepsize rule*

$$s_k = \gamma_k \frac{q^* - q(\mu^k)}{\|c^k\|^2}, \tag{17}$$

with $c^k \in \partial q(\mu^k)$ and $0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 2$, which was introduced by Poljak in [12] (see also Shor [13]).

Unfortunately, in most practical problems q^* and $q(\mu^k)$ are unknown. The latter can be approximated by $q_{\varepsilon_k}(\mu^k) = l(x^k, \mu^k)$ and q^* is replaced with an estimate q_{lev}^k . This leads to the stepsize rule

$$s_k = \gamma_k \frac{q_{lev}^k - q_{\varepsilon_k}(\mu^k)}{\|c^k\|^2}, \tag{18}$$

where $c^k \in \partial_{\varepsilon_k} q(\mu^k)$ is bounded for $k = 0, 1, \dots$

Dynamic Stepsize Rule with Adjustment Procedure (DSRAP). An option to estimate q^* is to use the *adjustment procedure* suggested by Nedić and Bertsekas [11], but fitted for the ε -subgradient method, its convergence was analyzed by Mijangos in [9] (see also [6]).

In this procedure q_{lev}^k is the best function value achieved up to the k th iteration, in our case $\max_{0 \leq j \leq k} q_{\varepsilon_j}(\mu^j)$, plus a positive amount δ_k , which is adjusted according to algorithm's progress.

The adjustment procedure obtains q_{lev}^k as follows:

$$q_{lev}^k = \max_{0 \leq j \leq k} q_{\varepsilon_j}(\mu^j) + \delta_k,$$

and δ_k is updated according to

$$\delta_{k+1} = \begin{cases} \rho\delta_k, & \text{if } q_{\varepsilon_{k+1}}(\mu^{k+1}) \geq q_{lev}^k, \\ \max\{\beta\delta_k, \delta\}, & \text{if } q_{\varepsilon_{k+1}}(\mu^{k+1}) < q_{lev}^k, \end{cases}$$

where δ_0, δ, β , and ρ are fixed positive constants with $\beta < 1$ and $\rho \geq 1$.

Dynamic Stepsize Rule with Relaxation-Level Control (DSRLC). Another choice to compute an estimate q_{lev}^k for (18) is to use a dynamic stepsize rule with relaxation-level control, which is based on the algorithm given by Brännlund [2], whose convergence was shown by Goffin and Kiwiel in [5] for $\varepsilon_k = 0$ for all k .

Mijangos in [9] fitted this method to the dual problem of NCNFP (4-5) for $\{\varepsilon_k\} \rightarrow 0$ and analyzed its convergence.

In this case, in contrast to the *adjustment procedure*, q^* is estimated by q_{lev}^k , which is a target level that is updated only if a sufficient ascent is detected or when the path long done from the last update exceeds a given upper bound B , see [9, 8, 6].

3 Solution to NCNFP

An algorithm is given below for solving **NCNFP**. This algorithm uses the approximate subgradient methods described in Section 2.

The value of the dual function $q(\mu^k)$ is estimated by minimizing approximately $l(x, \mu^k)$ over $x \in \mathcal{F}$ (the set defined by the network constraints) so that the optimality tolerance, τ_x^k , becomes more rigorous as k increases; i.e., the minimization will be *asymptotically exact* [1]. In other words, we set $q_{\varepsilon_k}(\mu^k) = l(x^k, \mu^k)$, where x^k minimizes approximately the nonlinear network subproblem **NNS_k**

$$\underset{x \in \mathcal{F}}{\text{minimize}} \quad l(x, \mu^k)$$

in the sense that this minimization stops when we obtain a x^k such that

$$\|Z^t \nabla_x l(x^k, \mu^k)\| \leq \tau_x^k$$

where $\lim_{k \rightarrow \infty} \tau_x^k = 0$ and Z represents the reduction matrix whose columns form a base of the null subspace of the subspace generated by the rows of the matrix of active network constraints of this subproblem.

Algorithm 1 (Approximate subgradient method for NCNFP)

Step 0. *Initialize.* Set $k = 1, N_{max}, \tau_x^1, \epsilon_\mu$ and τ_μ . Set $\mu^1 = 0$.

Step 1. *Compute* the dual function estimate, $q_{\epsilon_k}(\mu^k)$, by solving NNS_k , so that if $\|Z^t \nabla_x l(x^k, \mu^k)\| \leq \tau_x^k$, then $x^k \in \mathcal{F}$ is an approximate solution, $q_{\epsilon_k}(\mu^k) = l(x^k, \mu^k)$, and $c^k = c(x^k)$ is an ϵ_k -subgradient of q in μ^k .

Step 2. *Check the stopping rules* for μ^k .

$$T_1: \text{ Stop if } \max_{i=1, \dots, r} \left\{ \frac{(c_i^k)^+}{1 + (c_i^k)^+} \right\} < \tau_\mu, \text{ where } (c_i^k)^+ = \max\{0, c_i(x^k)\}.$$

$$T_2: \text{ Stop if } \left| \frac{q^k - (q^{k-1} + q^{k-2} + q^{k-3})/3}{1 + q^k} \right| < \epsilon_\mu, \text{ where } q^n = q_{\epsilon_n}(\mu^n).$$

$$T_3: \text{ Stop if } \frac{1}{4} \sum_{i=0}^3 \|\mu^{k-i} - \mu^{k-i-1}\|_\infty < \epsilon_\mu.$$

T_4 : Stop if k reaches a prefixed value N_{max} .

If μ^k fulfils one of these tests, then it is optimal, and the algorithm stops.

Without a duality gap, (x^k, μ^k) is a primal-dual solution.

Step 3. *Update* the estimate μ^k by means of the iteration

$$\mu_i^{k+1} = \begin{cases} \mu_i^k + s_k c_i^k, & \text{if } \mu_i^k + s_k c_i^k > 0 \\ 0, & \text{otherwise} \end{cases}$$

where s_k is computed using some stepsize rule. Go to Step 1.

In Step 0, for the checking of the stopping rules, $\tau_\mu = 10^{-5}$, $\epsilon_\mu = 10^{-5}$ and $N_{max} = 200$ have been taken. In addition, $\tau_x^0 = 10^{-1}$ by default.

Step 1 of this algorithm is carried out by the code PFNL, described in [10], which is based on the specific procedures for nonlinear networks flows [14] and active set procedures.

In Step 2, alternative heuristic tests have been used for practical purposes. T_1 checks the feasibility of x^k , as if it is feasible the duality gap is zero, and then (x^k, μ^k) is a primal-dual solution for **NCNFP**. T_2 and T_3 mean that μ does not improve for the last N iterations. Note that $N = 4$.

To obtain s_k in Step 3, we have alternately used the stepsize rules given in Section 2.1, that is, VCSR, DSR, DSRAP, and DSRLC.

The values given above have been heuristically chosen. The implementation in Fortran-77 of the previous algorithm, termed PFNRN05, was designed to solve large-scale nonlinear network flow problems with nonlinear side constraints.

4 Computational Results

The problems used in these tests were created by means of the following DIMACS-random-network generators: Rmfgen and Gridgen, see [4]. These generators provide linear flow problems in networks without side constraints. The inequality nonlinear side constraints for the DIMACS networks were generated

Table 1. Test problems

prob.	# arcs	# nodes	# side const.	# actives	# jac. entr.	# sb. arcs
c13e2	1524	360	360	10	364	89
c15e2	1524	360	180	59	307	116
c17e2	1524	360	216	82	367	132
c18e2	1524	360	360	152	615	201
c13n1	1524	360	360	38	364	681
c15n1	1524	360	180	110	307	620
c17n1	1524	360	216	129	367	615
c18n1	1524	360	360	205	615	588
c21e2	5420	1200	120	3	135	30
c22e2	5420	1200	600	13	656	46
c23e2	5420	1200	120	16	135	236
c24e2	5420	1200	600	91	1348	89
c33e2	4008	501	150	17	589	83
c34e2	4008	501	902	108	3645	106
c35e2	4008	501	251	28	1014	84
c38e2	4008	501	1253	135	4054	112
c42e2	12008	1501	15	4	182	174
c47e2	12008	1501	300	105	3617	327
c48e2	12008	1501	751	102	4537	236
c49e2	12008	1501	751	116	4537	258

Table 2. Computational results

Problem	VCSR			DSR			DSRAP			DSRLC		
	oit	$\ c\ $	t	oit	$\ c\ $	t	oit	$\ c\ $	t	oit	$\ c\ $	t
c13e2	24	10^5	1.2	8	10^3	0.7	7	10^8	0.7	7	10^9	0.6
c15e2	6	10^8	5.5	8	10^1	96.7	8	10^6	1.6	8	10^7	2.6
c17e2	7	10^7	10.7	1007	10^1	32.0	8	10^5	2.4	9	10^7	3.9
c18e2	11	10^{10}	33.1	664	10^1	43.6	11	10^7	39.1	8	10^6	8.3
c13n1	9	10^6	115.5	12	10^3	65.3	8	10^7	109.9	8	10^7	85.0
c15n1	10	10^6	82.4	18	10^3	188.4	8	10^6	63.7	21	10^6	157.0
c17n1	11	10^6	89.6	129	10^2	525.6	9	10^6	71.7	24	10^8	376.7
c18n1	11	10^6	123.1	10	10^8	370.8	10	10^8	332.5	44	10^9	1577.7
c21e2	24	10^6	1.5	28	10^2	1.5	7	10^7	1.6	7	10^7	1.6
c22e2	65	10^6	4.4	46	10^2	3.1	9	10^6	2.1	8	10^6	1.9
c23e2	31	10^5	6.5	47	10^1	6.4	10	10^7	6.9	11	10^7	6.4
c24e2	158	10^5	15.4	13	10^6	4.8	12	10^{10}	5.1	13	10^6	4.7
c33e2	7	10^6	1.5	10	10^3	1.5	8	10^7	1.6	8	10^6	1.7
c34e2	9	10^6	5.7	36	10^3	8.4	9	10^9	5.1	9	10^6	4.8
c35e2	8	10^6	2.3	11	10^3	2.1	9	10^9	2.3	9	10^8	2.5
c38e2	8	10^6	8.0	31	10^3	11.5	8	10^8	8.6	9	10^8	5.1
c42e2	10	10^7	11.9	10	10^3	18.8	7	10^8	14.3	8	10^{10}	14.2
c47e2	15	10^6	388.6	41	10^3	248.1	31	10^6	633.3	98	10^7	699.8
c48e2	11	10^6	122.5	43	10^3	98.1	11	10^8	235.0	48	10^6	138.3
c49e2	12	10^6	173.3	—	—	—	12	10^7	268.3	87	10^8	291.2

through the *Dirnl* random generator described in [10]. The last two letters indicate the type of objective function that we have used: Namur functions, **n1**, and EIO1 functions, **e2**.

The EIO1 family creates problems with a moderate number of superbasic variables (i.e., dimension of the null space) at the solution (# sb. arcs). By contrast, the Namur functions [14] generates a high number of superbasic arcs at the optimizer, see Table 1. More details about these problems can be found in [10].

In Table 2, for each method used to compute the stepsize, “oit” represents the number of outer iterations required and $\|c\|_\infty$ represents the maximum violation of the side constraints in the optimal solution; that is, it offers information about the feasibility of this solution and, hence, about its duality gap. The efficiency is evaluated by means of the run-times in CPU-seconds in the column *t*. The results point out that in spite of the simplicity of the variant of the constant step rule, VCSR, it is quite efficient and robust, in 8 of the 20 problems VCSR achieves the best CPU-times. However, DSRLC has a similar efficiency, but it is more accurate than VCSR (compare the columns $\|c\|_\infty$).

5 Conclusions

This paper has presented a variant of the constant step rule for approximate subgradient methods applied to solve nonlinearly constrained network flow problems. The convergence of the ε -subgradient method with the variant of the constant step rule has been analyzed. Moreover, its efficiency has been compared with that of other ε -subgradient methods. The results of the numerical tests encourage to carry out further experimentation, which also includes real problems, and to analyze more carefully the influence of some parameters over the performance of this code.

References

1. Bertsekas, D.P. *Nonlinear Programming*. 2nd ed., Athena Scientific, Belmont, Massachusetts (1999)
2. Brännlund, U. *On relaxation methods for nonsmooth convex optimization*. Doctoral Thesis, Royal Institute of Technology, Stockholm, Sweden (1993)
3. Correa, R. and Lemarechal, C. Convergence of some algorithms for convex minimization. *Mathematical Programming*, Vol. 62 (1993) 261–275
4. DIMACS. *The first DIMACS international algorithm implementation challenge : The bench-mark experiments*. Technical Report, DIMACS, New Brunswick, NJ, USA (1991)
5. Goffin, J.L. and Kiwiel, K. Convergence of a simple subgradient level method. *Mathematical Programming*, Vol. 85 (1999) 207–211
6. Kiwiel, K. Convergence of approximate and incremental subgradient methods for convex optimization. *SIAM J. on Optimization*, Vol. 14(3) (2004) 807–840
7. Mijangos, E. An efficient method for nonlinearly constrained networks. *European Journal of Operational Research*, Vol. 161(3) (2005) 618–635.

8. Mijangos, E. Efficient dual methods for nonlinearly constrained networks. Springer-Verlag Lecture Notes in Computer Science, Vol. 3483 (2005) 477–487
9. Mijangos, E. Approximate subgradient methods for nonlinearly constrained network flow problems, *Journal on Optimization Theory and Applications*, Vol. 128(1) (2006) (forthcoming)
10. Mijangos, E. and Nabona, N. The application of the multipliers method in nonlinear network flows with side constraints. Technical Report 96/10, Dept. of Statistics and Operations Research. Universitat Politècnica de Catalunya, 08028 Barcelona, Spain (1996) (downloadable from website <http://www.ehu.es/~mepmifee/>)
11. Nedić, A. and Bertsekas, D.P. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, Vol. 12(1) (2001) 109–138.
12. Poljak, B.T. Minimization of unsmooth functionals, *Z. Vyschisl. Mat. i Mat. Fiz.*, Vol. 9 (1969) 14–29
13. Shor, N.Z. Minimization methods for nondifferentiable functions, Springer-Verlag, Berlin (1985)
14. Toint, Ph.L. and Tuytens, D. On large scale nonlinear network optimization. *Mathematical Programming*, Vol. 48 (1990) 125–159

On the Optimal Buffer Allocation of an FMS with Finite In-Process Buffers

Soo-Tae Kwon

Department of Information Systems,
School of Information Technology and Engineering, Jeonju University

Abstract. This paper considers a buffer allocation problem of flexible manufacturing system composed of several parallel workstations each with both limited input and output buffers, where machine blocking is allowed and two automated guided vehicles are used for input and output material handling. Some interesting properties are derived that are useful for characterizing optimal allocation of buffers for the given FMS model. By using the properties, a solution algorithm is exploited to solve the optimal buffer allocation problem, and a variety of differently-sized decision parameters are numerically tested to show the efficiency of the algorithm.

Keywords: FMS, Queueing Network, Throughput, Buffer.

1 Introduction

Flexible manufacturing systems (FMSs) have been introduced in an effort to increase productivity by reducing inventory and increasing the utilization of machining centers simultaneously. An FMS combines the existing technology of NC manufacturing, automated material handling, and computer hardware and software to create an integrated system for the automatic random processing of palletized parts across various workstations in the system.

The design of an FMS begins with a survey of the manufacturing requirements of the products produced in the firm with a view to identifying the range of parts which should be produced on the FMS. Then the basic design concepts must be established. In particular the function, capability and number of each type of workstation, the type of material handling system and the type of storage should be determined.

At the detailed design stage it will be necessary to determine such aspects as the required accuracy of machines, tool changing systems and the method of feeding and locating parts at machines. Then the number of transport devices, the number of pallets, the capacity of central and local storages must be determined, along with some general strategies for work transport to reduce delays due to interference.

One of key questions that the designer face in an FMS is the buffer allocation problem, i.e., how much buffer storage to allow and where to place it within the system. This is an important question because buffers can have a great impact on

the efficiency of production. They compensate for the blocking and the starving of the workstations. Unfortunately, buffer storage is expensive both due to its direct cost and due to the increase of the work-in-process(WIP) inventories. Also, the requirement to limit the buffer storage can be a result of space limitations in the factory.

Much research has concentrated on queueing network model analyses to evaluate the performance of FMSs, and concerned with mathematical models to address the optimization problems of complex systems such as routing optimization, server allocation, workload allocation, buffer allocation on the basis of the performance model. Vinod and Solberg (1985) have presented a methodology to design the optimal system configuration of FMSs modeled as a closed queueing networks of multiserver queues. Buzacott and Yao (1986) have reviewed the work on modelling FMS with particular focus on analytical models. Shanthikumar and Yao (1989) have solved the optimal buffer allocation problem with increasing concave production profits and convex buffer space costs. Paradopoulos and Vidalis (1999) have investigated the optimal buffer allocation in short balanced production lines consisting of machines that are subject to breakdown. Enginarlar et al. (2002) have investigated the smallest level of buffering to ensure the desired production rate in serial lines with unreliable machines.

In the above-mentioned reference models, machines were assumed not to be blocked, that is, not to have any output capacity restriction. These days, the automated guided vehicle (AGV) is commonly used to increase potential flexibility. By the way, it may not be possible to carry immediately the finished parts from the machines which are subject to AGV's capacity restriction. The restriction can cause any operation blocking at the machines, so that it may be desirable to provide some storage space to reduce the impact of such blocking. In view of reducing work-in-process storage, it is also required to have some local buffers of proper size at each workstation. Sung and Kwon(1994) have investigated a queueing network model for an FMS composed of several parallel workstations each with both limited input and output buffers where two AGVs are used for input and output material handling, and Kwon (2005) has considered a workload allocation problem on the basis of the same model.

In this paper, a buffer allocation problem is considered to yield the highest throughput for the given FMS model (Sung and Kwon 1994). Some interesting properties are derived that are useful for characterizing optimal allocation of buffers, and some numerical results are presented.

2 The Performance Evaluation Model

The FMS model is identical to that in Sung and Kwon(1994). The network consists of a set of n workstations. Each workstation i ($i = 1, \dots, n$) has machines with both limited input and output buffers. The capacities of input and output buffers are limited up to IB_i and OB_i respectively, and the machines perform in an exponential service time distribution. All the workstations are linked to an automated storage and retrieval system (AS/RS) by AGVs which consist of

AGV(I) and AGV(O). The capacity of the AS/RS is unlimited, and external arrivals at the AS/RS follow a Poisson process with rate λ .

The FCFS (first come first served) discipline is adopted here for the services of AGVs and machines. AGV(I) delivers the input parts from the AS/RS to each input buffer of workstations, and AGV(O) carries the finished parts away from each output buffer of workstations to the AS/RS, with corresponding exponential service time distributions. Specifically, AGV(I) distributes all parts from the AS/RS to the workstations according to the routing probabilities γ_i ($\sum_{i=1}^n \gamma_i = 1$) which can be interpreted as the proportion of part dispatching from the AS/RS to workstation i .

Moreover, any part (material) can be blocked on arrival (delivery) at an input buffer which is already full with earlier-arrived parts. Such a blocked part will be recirculated instead of occupying the AGV(I) and waiting in front of the workstation (block-and-recirculate mechanism). Any finished part can also be blocked on arrival at an output buffer which is already full with earlier-finished parts. Such a blocked part will occupy the machine to remain blocked until a part departure occurs from the output buffer. During such a blocking time, the machine cannot render service to any other part that might be waiting at its input buffer (block-and-hold mechanism).

Sung and Kwon(1994) have developed an iterative algorithm to approximate system performance measures such as system throughput and machine utilization. The approximation procedure decomposes the queueing network into individual queues with revised arrival and service processes. These individual queues are then analyzed in isolation. The individual queues are grouped into two classes. The first class consists of input buffer and machine, and the second one consists of output buffers and AGV(O). The first and second classes are called the first-level queue and the second-level queue, respectively.

The following notations are used throughout this paper ($i = 1, \dots, n$):

- λ external arrival rate at AS/RS
- λ_i arrival rate at each input buffer i in the first-level queue
- λ_i^* arrival rate at each output buffer i in the second-level queue
- μ service rate of AGV
- μ_i service rate of machine i
- $P(k_1, \dots, k_n)$ probability that there are k_i units at each output buffer i in the second-level queue with infinite capacity.
- $P(idle)$ probability that there is no unit in the second-level queue with infinite capacity.
- $\prod(k_1, \dots, k_n)$ probability that there are k_i units at each output buffer i in the second-level queue with finite capacity.
- $\prod(idle)$ probability that there is no unit in the second-level queue with finite capacity.

The second-level queue is independently analyzed first to find the steady-state probability by using the theory of reversibility. The steady-state probability is derived as follows.

Lemma 1. (refer to Sung and Kwon 1994, Theorem 2)

The steady-state probability of the second-level queue is derived as

$$\prod (k_1, \dots, k_n) = P(k_1, \dots, k_n)/G$$

$$\prod (idle) = P(idle)/G \tag{1}$$

where,

$$A = \{(k_1, \dots, k_n) | 0 \leq k_i \leq OB_i, 1 \leq i \leq n\},$$

$$G = \sum_{(k_1, \dots, k_n) \in A} P(k_1, \dots, k_n) + P(idle),$$

$$P(k_1, \dots, k_n) = (1 - \rho) \cdot \rho^{(k_1 + \dots + k_n + 1)} \cdot \frac{(k_1 + \dots + k_n)!}{k_1! \dots k_n!} \cdot q_1^{k_1} \dots q_n^{k_n},$$

$$P(idle) = 1 - \rho,$$

$$\rho = \sum_{i=1}^n \lambda_i^* / \mu,$$

$$q_i = \lambda_i^* / \sum_{i=1}^n \lambda_i^*.$$

It is followed by finding the clearance service time accommodating all the possible blocking delays that a part might undergo due to the phenomenon of blocking. The clearance time is derived from the steady-state probability of second-level queue. Then, the first-level queues are analyzed by this expected clearance time in the approach of the M/M/1/K queueing model.

3 The Buffer Allocation Problem

In FMS, a frequently encountered problem is concerned with how to allocate buffer space among several subsystems for maximizing the production rate (system throughput). In this section, the buffer allocation problem is considered to yield the highest throughput for the given performance evaluation model. The optimal buffer allocation problem can be stated as follows :

$$\begin{aligned} \text{Maximize} \quad & Z = TH(x_1, \dots, x_n, x_{n+1}, \dots, x_{2n}) \\ \text{s.t.} \quad & \sum_{i=1}^{2n} x_i \leq S \end{aligned} \tag{2}$$

where

- $TH(x_1, \dots, x_n, x_{n+1}, \dots, x_{2n})$ = the system throughput,
- x_i = the number of buffers allocated to buffer i (input buffer : $1 \leq i \leq n$,
output buffer : $n + 1 \leq i \leq 2n$), that is $(IB_1, \dots, IB_n, OB_1, \dots, OB_n)$
- S = the maximum total number of buffers to be allocated.

Despite of its practical importance, this buffer allocation problem has not been successfully studied in the literature. The major difficulty appears to be lack of known properties regarding the throughput of system as a function of its buffer capacity. Some interesting properties for the associated system throughput are now derived.

Property 1. In the first-level queue-alone subsystem, the throughput is a monotonically increasing concave function of its buffer size.

Proof:

Let $\rho_i (= \lambda_i / \mu_i)$ and IB_i denote the utilization and the buffer size of the first-level queue i , respectively. Then the throughput of the first-level queue i , $TH(\lambda_i, IB_i, \mu_i)$, can be derived as follows :

$$TH(\lambda_i, IB_i, \mu_i) = \mu_i \cdot \left(1 - \frac{1 - \rho_i}{1 - \rho_i^{IB_i+1}} \right)$$

By the definition of $TH(IB_i)$,

$$\begin{aligned} & TH(\lambda_i, IB_i + 1, \mu_i) - TH(\lambda_i, IB_i, \mu_i) \\ &= \mu_i \cdot \left(\frac{1 - \rho_i}{1 - \rho_i^{IB_i+1}} - \frac{1 - \rho_i}{1 - \rho_i^{IB_i+2}} \right) \\ &= \mu_i \cdot \left(\frac{1}{1 + \rho_i + \rho_i^2 + \dots + \rho_i^{IB_i}} - \frac{1}{1 + \rho_i + \rho_i^2 + \dots + \rho_i^{IB_i+1}} \right) \\ &> 0 \quad \text{for all } \rho_i. \end{aligned}$$

And,

$$\begin{aligned} & 2TH(\lambda_i, IB_i + 1, \mu_i) - TH(\lambda_i, IB_i, \mu_i) - TH(\lambda_i, IB + 2, \mu_i) \\ &= 2 \cdot \mu_i \cdot \left(1 - \frac{1 - \rho_i}{1 - \rho_i^{IB_i+2}} \right) - \mu_i \cdot \left(1 - \frac{1 - \rho_i}{1 - \rho_i^{IB_i+1}} \right) - \mu_i \cdot \left(1 - \frac{1 - \rho_i}{1 - \rho_i^{IB_i+3}} \right) \\ &= \mu_i \cdot (1 - \rho_i) \left[\frac{1}{1 - \rho_i^{IB_i+1}} + \frac{1}{1 - \rho_i^{IB_i+3}} - \frac{2}{1 - \rho_i^{IB_i+2}} \right] \\ &= \frac{\mu_i \cdot (1 - \rho_i)}{(1 - \rho_i^{IB_i+1})(1 - \rho_i^{IB_i+3})(1 - \rho_i^{IB_i+2})} \cdot [\rho_i^{IB_i+3} + \rho_i^{IB_i+1} - 2\rho_i^{IB_i+2} \\ &\quad + \rho_i^{2 \cdot IB_i+5} + \rho_i^{2 \cdot IB_i+3} - 2\rho_i^{2 \cdot IB_i+4}] \\ &= \frac{\mu_i \cdot (1 - \rho_i)^3 \cdot (\rho_i^{IB_i+1} + \rho_i^{2 \cdot IB_i+3})}{(1 - \rho_i^{IB_i+1})(1 - \rho_i^{IB_i+3})(1 - \rho_i^{IB_i+2})} \\ &> 0 \quad \text{for all } \rho_i. \end{aligned}$$

Thus, the throughput of the first-level queue is a monotonically increasing concave function of buffer size. This completes the proof.

Also, the throughput of the second-level queue is characterized as follows.

Property 2. In the second-level queue-alone subsystem, the throughput is a monotonically increasing concave function of its buffer size.

Proof:

Let $\rho (= \sum_{i=1}^n \frac{\lambda_i}{\mu})$ and (OB_1, \dots, OB_n) denote the utilization and the output buffer sizes of the second-level queue, respectively. Then the throughput of the second-level queue, $TH(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_n, \mu)$, can be derived as follows.

$$\begin{aligned} TH(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_n, \mu) &= \mu \cdot \left(1 - \prod (idle) \right) \\ &= \mu \cdot \left(1 - \frac{1 - \rho}{G(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_n, \mu)} \right) \\ &= \mu \cdot \left(1 - \frac{1}{\phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_n, \mu)} \right) \end{aligned}$$

where,

$$\begin{aligned}
 G(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_n, \mu) &= (1 - \rho) \left[1 + \sum_{k_1=0}^{OB_1} \dots \sum_{k_n=0}^{OB_n} \rho^{n+1} \frac{(k_1 + \dots + k_n)!}{k_1! \dots k_n!} q_1^{k_1} \dots q_n^{k_n} \right], \\
 \phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_n, \mu) &= 1 + \sum_{k_1=0}^{OB_1} \dots \sum_{k_n=0}^{OB_n} \rho^{n+1} \frac{(k_1 + \dots + k_n)!}{k_1! \dots k_n!} q_1^{k_1} \dots q_n^{k_n} \\
 n &= k_1 + \dots + k_n
 \end{aligned}$$

And, let

$$\begin{aligned}
 \psi_1 &= \sum_{k_1=0}^{OB_1} \dots \sum_{k_i=OB_i+1}^{OB_i+1} \dots \sum_{k_n=0}^{OB_n} \rho^{n+1} \frac{(k_1 + \dots + k_n)!}{k_1! \dots k_n!} q_1^{k_1} \dots q_n^{k_n}, \quad \text{and} \\
 \psi_2 &= \sum_{k_1=0}^{OB_1} \dots \sum_{k_i=OB_i+2}^{OB_i+2} \dots \sum_{k_n=0}^{OB_n} \rho^{n+1} \frac{(k_1 + \dots + k_n)!}{k_1! \dots k_n!} q_1^{k_1} \dots q_n^{k_n}
 \end{aligned}$$

These lead to the relation $\psi_1 > \psi_2$, and it holds that

$$\begin{aligned}
 \phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i + 1, \dots, OB_n, \mu) &= \phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i, \dots, OB_n, \mu) + \psi_1 \\
 \phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i + 2, \dots, OB_n, \mu) &= \phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i, \dots, OB_n, \mu) + \psi_1 + \psi_2
 \end{aligned}$$

By the definition of $TH(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_n, \mu)$

$$\begin{aligned}
 TH(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i + 1, \dots, OB_n, \mu) &- TH(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i, \dots, OB_n, \mu) \\
 = \mu \cdot \left[\frac{1}{\phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i, \dots, OB_n, \mu)} \right. & \\
 \left. - \frac{1}{\phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i, \dots, OB_n, \mu) + \psi_1} \right] & \\
 > 0 \quad \text{for all } OB_i &
 \end{aligned}$$

And, $2 \cdot TH(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i + 1, \dots, OB_n, \mu)$

$$\begin{aligned}
 - TH(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i, \dots, OB_n, \mu) &- TH(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i + 2, \dots, OB_n, \mu) \\
 = \mu \cdot \left[\frac{-2}{\phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i, \dots, OB_n, \mu) + \psi_1} \right. & \\
 \left. + \frac{1}{\phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i, \dots, OB_n, \mu)} \right] &
 \end{aligned}$$

$$\begin{aligned}
 & + \frac{1}{\phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i, \dots, OB_n, \mu) + \psi_1 + \psi_2} \\
 = & \mu \cdot [\phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_n, \mu) \cdot (\psi_1 - \psi_2) + \psi_1^2 + \psi_1 \cdot \psi_2] \\
 & \cdot \frac{1}{\phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i, \dots, OB_n, \mu) + \psi_1} \\
 & \cdot \frac{1}{\phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i, \dots, OB_n, \mu)} \\
 & \cdot \frac{1}{\phi(\lambda_1^*, \dots, \lambda_n^*, OB_1, \dots, OB_i, \dots, OB_n, \mu) + \psi_1 + \psi_2} \\
 > 0 \quad \text{for all } OB_i
 \end{aligned}$$

Thus, the throughput of the second-level queue is a monotonically increasing concave function of buffer size. This completes the proof.

Finally, the following result can be obtained by the comparison of the throughputs for buffer allocation scheme.

Property 3. In the second-level queue-alone subsystem, the balanced buffer allocation scheme maximizes the throughput.

Proof:

For simplification, the proof will be completed only for the case of $n = 2, S = 2$ and $\lambda_1 = \lambda_2$. Let TH^b and TH^{nb} be the throughput of the balanced allocation scheme case ($OB_1 = OB_2 = 1$) and the other one ($OB_1 = 2$), respectively.

$$TH^b - TH^{nb} = \mu \left(1 - \frac{1 - \rho}{G^b} \right) - \mu \left(1 - \frac{1 - \rho}{G^{nb}} \right) = \mu(1 - \rho) \frac{G^b - G^{nb}}{G^b \cdot G^{nb}}$$

Since $q_1 = q_2$ and by the definition of G ,

$$\begin{aligned}
 G^b - G^{nb} &= (1 - \rho)[(1 - \rho + \rho^2 + 2\rho^3 q_1 q_2) - (1 + \rho + \rho^2 q_1 + \rho^3 q_1^2)] \\
 &= (1 - \rho)[\rho^2(1 - q_1) + \rho^3(2q_1 q_2 - q_1^2)] \\
 &= (1 - \rho)\left[\frac{1}{2}\rho^2 + \frac{1}{2}\rho^3\right]
 \end{aligned}$$

Therefore, $TH^b - TH^{nb} \geq 0$. This complete the proof.

On the basis of properties, now consider the optimization problem in (2). Since the throughput is a monotonically increasing concave function of buffer capacities for both first-level and second-level queue as proved, the marginal allocation approach of Fox (1966) can be used to efficiently solve the optimal buffer allocation problem in (2). The idea of this approach is as follows.

Let $\Delta TH(x_1, \dots, x_i, \dots, x_{2n}) = TH(x_1, \dots, x_i + 1, \dots, x_{2n}) - TH(x_1, \dots, x_i, \dots, x_{2n})$ for all i . Allocate the available buffer spaces to the buffer that would yield the largest increase in $\Delta F(x_i)$ one at a time. Continue this procedure until the available buffer spaces are exhausted.

The algorithm of the buffer allocation problem can be summarized as follows.

- Step 1. Set $x_i = 0$, for all $i(= 1, \dots, 2n)$.
- Step 2. For all i , calculate $\Delta TH(x_1, \dots, x_i, \dots, x_{2n}) = TH(x_1, \dots, x_i + 1, \dots, x_{2n}) - TH(x_1, \dots, x_i, \dots, x_{2n})$ by using the performance evaluation model.

Table 1. The result of the parameter set 1

iteration	$x_1 = IB_1$	$x_2 = OB_2$	throughput	increment	allocation
0	0	0	.5262	-	
1	1	0	.6667	.1405	**
	0	1	.5846	.4941	
2	2	0	.7313	.0646	**
	1	1	.7311	.0644	
3	3	0	.7677	.0364	
	2	1	.7965	.0652	**
4	3	1	.83288	.03638	**
	2	2	.83287	.03637	
5	4	1	.8558	.0229	
	3	2	.8685	.0356	**

Table 2. The result of the parameter set 2

S	x_1	x_2	x_3	x_4	TH	Δ	allocation	S	x_1	x_2	x_3	x_4	TH	Δ	allocation
	IB_1	IB_2	IB_3	IB_4					IB_1	IB_2	IB_3	IB_4			
0	0	0	0	0	.6291	-									
1	1	0	0	0	.7043	.0752	**	6	3	1	1	1	.8867	.0142	
	0	1	0	0	.7043	.0752		2	2	1	1	1	.9003	.0278	**
	0	0	1	0	.6641	.035		2	1	2	1	1	.882	.0095	
	0	0	0	1	.6641	.035		2	1	1	2	2	.89	.0175	
2	2	0	0	0	.7378	.033		7	3	2	1	1	.9149	.0146	**
	1	1	0	0	.7758	.0715	**	2	3	1	1	1	.9149	.0146	
	1	0	1	0	.731	.0267		2	2	2	1	1	.912	.0117	
	1	0	0	1	.7438	.0395		2	2	1	2	2	.912	.0117	
3	2	1	0	0	.8087	.0329	**	8	4	2	1	1	.9232	.0083	
	1	2	0	0	.8087	.0329		3	3	1	1	1	.9293	.0144	**
	1	1	1	0	.807	.0312		3	2	2	1	1	.9217	.0068	
	1	1	0	1	.807	.0312		3	2	1	2	2	.9284	.0135	
4	3	1	0	0	.8268	.0181		9	4	3	1	1	.9376	.0083	
	2	2	0	0	.8406	.0319		3	4	1	1	1	.9376	.0083	
	2	1	1	0	.8312	.0225		3	3	2	1	1	.9379	.0086	**
	2	1	0	1	.8425	.0338	**	3	3	1	2	2	.9379	.0086	
5	3	1	0	1	.8629	.0204		10	4	3	2	1	.9427	.0048	
	2	2	0	1	.8663	.0238		3	4	2	1	1	.9479	.01	
	2	1	1	1	.8725	.03	**	3	3	3	1	1	.9394	.0015	
	2	1	0	2	.855	.0125		3	3	2	2	2	.9524	.0145	**

- Step 3.** Find k such that $\Delta TH(x_1, \dots, x_k, \dots, x_{2n})$

$$= \underset{1 \leq i \leq 2n}{Max} \Delta TH(x_1, \dots, x_i, \dots, x_{2n}).$$
 Set $x_k = x_k + 1, S = S - 1$.
 If $S > 0$, then go to step 2.
- Step 4.** Stop.

In order to illustrate the solution procedure, the buffer allocation problem on the basis of the given performance evaluation model is considered with parameter set 1 ($\lambda = 1, \mu_1 = \mu_2 = 2, \mu = 1, S = 5$), parameter set 2 ($\lambda = 1, r_1 = r_2 = 0.5, \mu_1 = \mu_2 = 2, \mu = 1, S = 10$) and parameter set 3 ($\lambda = 1, r_1 = r_2 = 0.5, \mu_1 = 1, \mu_2 = 2, \mu = 1, S = 10$), where S, λ, r_i, μ_i , and μ denote the maximum total number of buffers to be allocated, the arrival rate, the routing probability, the machine service rate, and the AGV service rate, respectively.

At first, a simple system with a single workstation is considered to illustrate the marginal allocation procedure, which is identical to the two stage transfer line. The results of the system with parameter set 1 are shown in Table 1. The table gives both the amount of increment and throughput results.

At second, the buffer allocation problem is considered to test the efficiency of the solution algorithm. For $S = 1, \dots, 10$, the results of the system with parameter set 2 and 3 are shown in Table 2 and 3, respectively.

Table 3. The result of the parameter set 3

S	x_1	x_2	x_3	x_4	TH	Δ	allocation	S	x_1	x_2	x_3	x_4	TH	Δ	allocation
	IB_1	IB_2	IB_3	IB_4					IB_1	IB_2	IB_3	IB_4			
0	0	0	0	0	.5944	-									
1	1	0	0	0	.6662	.0718		6	3	2	1	0	.8547	.0162	
	0	1	0	0	.6729	.0785	**		2	3	1	0	.8592	.0207	
	0	0	1	0	.6219	.0275			2	2	2	0	.8502	.0117	
	0	0	0	1	.6297	.0353			2	2	1	1	.8692	.0307	**
2	1	1	0	0	.741	.0681	**	7	3	2	1	1	.8869	.0177	**
	0	2	0	0	.7074	.0345			2	3	1	1	.8835	.0143	
	0	1	1	0	.7049	.032			2	2	2	1	.8858	.0166	
	0	1	0	1	.7002	.0273			2	2	1	2	.8789	.0097	
3	2	1	0	0	.7731	.0321		8	4	2	1	1	.898	.0111	
	1	2	0	0	.7745	.0335	**		3	3	1	1	.901	.0141	
	1	1	1	0	.7728	.0318			3	2	2	1	.9014	.0145	**
	1	1	0	1	.771	.03			3	2	1	2	.8973	.0104	
4	2	2	0	0	.8061	.0316		9	4	2	2	1	.9105	.0091	
	1	3	0	0	.7931	.0186			3	3	2	1	.9177	.0163	**
	1	2	1	0	.8095	.035	**		3	2	3	1	.9068	.0054	
	1	2	0	1	.796	.0215			3	2	2	2	.9162	.0148	
5	2	2	1	0	.8385	.029	**	10	4	3	2	1	.9269	.0092	
	1	3	1	0	.8303	.0208			3	4	2	1	.9275	.0098	
	1	2	2	0	.8229	.0134			3	3	3	1	.924	.0063	
	1	2	1	1	.8376	.0281			3	3	2	2	.9281	.0104	**

The computational results present that the solution algorithm is very efficient. In case of $S = s$, the solution algorithm generated the optimal solution at the s -th iteration. However, it is impossible in practice to allocate buffer spaces by conventional approach, since the number of allocating combination becomes explosively large as the number of S and workstation increase.

And, the results of Table 2 and 3 imply that the balanced buffer allocation scheme maximize the system throughput. That is, in order to maximize the system throughput, the buffer should be allocated depending on the routing probability and service rate of machine and AGV.

4 Conclusions

In this paper, a design aspect of a flexible manufacturing system composed of several parallel workstations each with both limited input and output buffers where two AGVs are used for input and output material handling is considered. The optimal design decision is made on the allocation of buffer spaces on the basis of the given performance evaluation model.

Some interesting properties are derived that are useful for characterizing optimal allocation of buffer spaces. The properties are then used to exploit a solution algorithm for allocating buffer spaces. A variety of differently-sized decision parameters are numerically tested to show the efficiency of the algorithm. The results present that the solution algorithm is very efficient.

Further research is to consider the cost factor more explicitly, and also to extend these concepts to general production systems.

References

1. Buzacott, J.A. and Yao, D.D.: Flexible Manufacturing Systems: A Review of Analytical Models. *Management Science*, **32**, No. 7, (1986) 890-905.
2. Enginarlar, E., Li, J., Meerkov, S.M., and Zhang, R.Q.: Buffer Capacity for Accommodating Machine Downtime in Serial Production Lines, *International Journal of Production Research*, **40**, No. 3, (2002) 601-624.
3. Fox, B.: Discrete Optimization via Marginal Analysis, *Management Science*, **13**, No. 3, (1966) 210-216.
4. Kwon, S.T.: On the Optimal Workloads Allocation of an FMS with Finite In-process Buffers, *Lecture Notes in Computer Science*, **3483**, (2005) 624-631.
5. Papadopoulos, H.T. and Vidalis, M.I.: Optimal Buffer Allocation in short-balanced unreliable production lines, *Computers & Industrial Engineering*, **37**, (1999) 691-710.
6. Shanthikumar, J. G. and Yao, D. D.: Optimal Buffer Allocation in a Multicell System, *The International Journal of Flexible Manufacturing Systems*, **1**, (1989) 347-356.
7. Sung, C.S. and Kwon, S.T.: Performance Modelling of an FMS with Finite Input and Output Buffers. *International Journal of Production Economics*, **37**, (1994) 161-175.
8. Vinod, B. and Solberg, J.J.: The Optimal Design of Flexible Manufacturing Systems, *International Journal of Production Research*, **23**, No. 6, (1985) 1141-1151.

Optimization Problems in the Simulation of Multifactor Portfolio Credit Risk

Wanmo Kang¹ and Kyungsik Lee^{2,*,**}

¹ Columbia University, New York, NY, USA

² Hankuk University of Foreign Studies, Yongin, Korea

Abstract. We consider some optimization problems arising in an efficient simulation method for the measurement of the tail of portfolio credit risk. When we apply an importance sampling (IS) technique, it is necessary to characterize the important regions. In this paper, we consider the computation of directions for the IS, which becomes hard in multifactor case. We show this problem is NP-hard. To overcome this difficulty, we transform the original problem to subset sum and quadratic optimization problems. We support numerically that these reformulation is computationally tractable.

1 Introduction

Measurement of portfolio credit risk is an important problem in financial industry. To reserve economic capital or to summarize the potential risk of a company, the portfolio credit risk is calculated frequently. Some of key properties of this measurement are the importance of dependence structure of obligors constituting the portfolio and the rare-event characteristic of large losses. Dependence among obligors incurs large losses more frequently, even though they are still rare. Gaussian copula is one of the most popular correlation structure in practice. (See [5].) Since there is no known analytical or numerical way to compute the tail losses under Gaussian copula framework, Monte Carlo method is a viable way to accomplish this task. (See [1], [4], [6], [8], and [9]) However, the rareness of large losses makes a crude Monte Carlo method impractical. To accelerate the simulation, one effective way is the application of IS. When applying IS, the identification of important region is the key for the efficiency enhancement. In this paper, we consider the problem of identifying important regions. The combinatorial complexity underlying this problem makes it a hard problem. We re-formulate this problem as a combination of quadratic optimizations and subset sum problems. The worst case complexity is not reduced, but the subset sum problems can be solved very fast in practice. Consequently this new approach works very well for actual problem instances.

* Corresponding author.

** This work was supported by Hankuk University of Foreign Studies Research Fund.

2 Portfolio Credit Risk and Importance Sampling

We briefly introduce the portfolio credit risk model and Gaussian copula framework. We consider the distribution of losses from defaults over a fixed horizon. We are interested in the estimation of the probability that the credit loss of a portfolio exceeds a given threshold. As it is difficult to estimate correlations among the default events of obligors, latent variables are introduced as default triggers and the dependence structure is imposed on the latent variables indirectly. A linear factor model is adopted for the correlations among them. We use the following notation:

- m : the number of obligors to which the portfolio is exposed;
- Y_k : default indicator (= 1 for default, = 0 otherwise) for the k -th obligor;
- p_k : marginal probability that the k -th obligor defaults;
- c_k : loss resulting from default of the k -th obligor;
- $L_m = c_1Y_1 + \dots + c_mY_m$: total loss from defaults.

We are interested in the estimation of $P(L_m > x)$ for a given threshold x when the event $\{L_m > x\}$ is rare. We call such one as a *large loss* event. We introduce latent normal random variables X_k for each Y_k . X_k 's are standard normal random variables and We set $Y_k = \mathbf{1}\{X_k > \Phi^{-1}(1 - p_k)\}$, with Φ the cumulative normal distribution. For a linear factor representation of X_k , we assume the following: There are d factors and t types of obligors. $\{\mathcal{I}_1, \dots, \mathcal{I}_t\}$ is a partition of the set of obligors $\{1, \dots, m\}$ into types. If $k \in \mathcal{I}_j$, then the k -th obligor is of type j and its latent variable is given by

$$X_k = \mathbf{a}_j^\top \mathbf{Z} + b_j \varepsilon_k$$

where $\mathbf{a}_j \in \mathbb{R}^d$ with $0 < \|\mathbf{a}_j\| < 1$, \mathbf{Z} is a d dimensional standard normal random vector, $b_j = \sqrt{1 - \mathbf{a}_j^\top \mathbf{a}_j}$ and ε_k are independent standard normal random variables. This dependence structure is called a *Gaussian copula model*. (See [2].) \mathbf{Z} represents common systematic risk factors and ε_k an idiosyncratic risk factor. We set $x = q \sum_{k=1}^m c_k$ for a given q , $0 < q < 1$. Denote the average loss of each type by $C_j = \sum_{k \in \mathcal{I}_j} c_k / |\mathcal{I}_j|$ and total average loss by $C = \sum_{k=1}^m c_k / m$. Then index sets (sets of types) important for the large losses exceeding qC can be characterized by the following index set $\mathcal{J} \subset \{1, \dots, t\}$ (See [3]):

$$\max_{\substack{\mathcal{J} \\ \neq \emptyset}} \sum_{j \in \mathcal{J}} C_j < qC \leq \sum_{j \in \mathcal{J}} C_j. \tag{1}$$

This characterization of an important index set can be interpreted as follows: to observe samples with large losses, the common factors should have values which enable the sum of average losses of the types in the index set to exceed the loss threshold.

After identifying these index sets (say \mathcal{J} , the point to shift the mean vectors of Gaussian common factors is found by

$$\boldsymbol{\mu}_{\mathcal{J}} := \operatorname{argmin} \{\|\mathbf{z}\| : \mathbf{z} \in G_{\mathcal{J}}\} \tag{2}$$

where

$$G_j := \{ \mathbf{z} \in \mathbb{R}^d : \mathbf{a}_j^\top \mathbf{z} \geq d_j \} \quad \text{and} \quad G_{\mathcal{J}} := \bigcap_{j \in \mathcal{J}} G_j .$$

$d_j > 0$ is a constant for each type calculated from the problem instance. In this paper, the positivity of d_j is sufficient.

Now returning to the Monte Carlo simulation, we sample the common factors from the mixture distribution of $N(\boldsymbol{\mu}_{\mathcal{J}}, \mathbf{I})$ for all the \mathcal{J} 's satisfying (1). $\boldsymbol{\mu}_{\mathcal{J}}$ is the minimum distant point to the important region for the large losses and we sample from the normal distribution shifted to those points. As usual, we compensate this change of measure by multiplying likelihood ratios. (See [3] for details.)

Define \mathcal{S}_q be the set of index sets satisfying (1). In principle, we can use $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$ in the simulation as the mean vectors of mixture distribution if $K = |\mathcal{S}_q|$. However, K becomes very large as t increases. The size depends on q , but for q values near 0.5, the order of K follows exponential to t . So identifying \mathcal{S}_q first and then finding corresponding $\boldsymbol{\mu}_{\mathcal{J}}$'s are impractical. In the next section, we exploit some structural properties and re-formulate the problem into a tractable one.

3 Re-formulation of Problem

The first idea comes from the fact that we use $\boldsymbol{\mu}_{\mathcal{J}}$ for the shift of mean vectors but \mathcal{J} is not explicitly used. So if $\boldsymbol{\mu}_{\mathcal{J}} = \boldsymbol{\mu}_{\mathcal{J}'}$ for two different index sets \mathcal{J} and \mathcal{J}' , we don't need to know what the two index sets are, but just need $\boldsymbol{\mu}_{\mathcal{J}}$. Hence we focus on the characterization of

$$\mathcal{V} := \{ \boldsymbol{\mu}_{\mathcal{J}} : \mathcal{J} \subset \{1, \dots, t\}, |\mathcal{J}| \leq d, G_{\mathcal{J}} \neq \emptyset \}$$

instead of \mathcal{S}_q which possibly consists of exponentially many elements with respect to the number of types t . The issue here is how to find the candidate IS distributions as fast as possible when the number of types, t , and the dimension of factors, d , are fixed.

3.1 Reduction of Candidate Mean Vectors

For a given problem instance, the size of \mathcal{S}_q depends on the value q . In the worst case, the size of \mathcal{S}_q will be $\binom{t}{\lfloor t/2 \rfloor}$, in which case the application of IS is intractable for instances with a large number of types. To avoid this difficulty, we need to devise a method that does not involve an explicit enumeration of the index sets in \mathcal{S}_q . The key fact is the following lemma.

Lemma 1. *For any $\mathcal{J} \in \mathcal{S}_q$ satisfying $G_{\mathcal{J}} \neq \emptyset$, there exists a $\mathcal{J}' \subset \mathcal{J}$ with $|\mathcal{J}'| \leq d$ such that*

$$\boldsymbol{\mu}_{\mathcal{J}} = \boldsymbol{\mu}_{\mathcal{J}'} .$$

Proof. (Sketch of Proof) Recall that the definition (2) implies that $\boldsymbol{\mu}_{\mathcal{J}}$ is the optimal solution of linear programming (LP), $\min\{\boldsymbol{\mu}_{\mathcal{J}}^{\top} \mathbf{z} : \mathbf{a}_j^{\top} \mathbf{z} \geq d_j \text{ for } j \in \mathcal{J}\}$. The LP duality gives a dual optimal solution π with $\mathcal{P} := \{j : \pi_j > 0\}$ and $|\mathcal{P}| \leq d$. The complementary slackness condition shows that $\boldsymbol{\mu}_{\mathcal{J}}$ and $\frac{\pi_j}{\|\mathbf{v}_j\|}$, $j \in \mathcal{P}$ satisfy KKT optimality conditions for $\min\{\|\mathbf{z}\| : \mathbf{a}_j^{\top} \mathbf{z} \geq d_j \text{ for } j \in \mathcal{P}\}$ and its dual. We can take $\mathcal{J}' = \mathcal{P}$ and this completes the proof. Refer [3] for details. \square

The lemma tells us that we don't have to spend our effort to solve (2) if $|\mathcal{J}| > d$. This gives a large reduction of our search space. From this lemma, we also have the following upper bound on the number of (2) which we have to solve.

Lemma 2. *For an instance with d factors and t types,*

$$|\{\boldsymbol{\mu}_{\mathcal{J}} : \mathcal{J} \in \mathcal{S}_q\}| \leq \binom{t}{d} + \binom{t}{d-1} + \dots + t < t^d.$$

Proof. Note that the righthand side of inequality is the number of ways of choosing d or less constraints from t candidates. Combining with Lemma 1, we complete the proof. \square

3.2 Derivation of the Subset Sum Problem

Recall that $\{\boldsymbol{\mu}_{\mathcal{J}} : \mathcal{J} \in \mathcal{S}_q\} \subset \mathcal{V}$ from Lemma 1. The upper bound in Lemma 2 is also an upper bound on $|\mathcal{V}|$. Our approach is to find \mathcal{V} , as reduced candidate mean vectors, and use it to get $\{\boldsymbol{\mu}_{\mathcal{J}} : \mathcal{J} \in \mathcal{S}_q\}$ from \mathcal{V} . Assume a representation $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ and define $\mathcal{H}(\mathbf{v}) := \{j : \mathbf{a}_j^{\top} \mathbf{v} \geq d_j, j = 1, \dots, t\}$ for $\mathbf{v} \in \mathcal{V}$. $\mathcal{H}(\mathbf{v})$ is the maximal index set satisfying $\mathbf{v} = \boldsymbol{\mu}_{\mathcal{H}(\mathbf{v})}$. Consider, for each $\mathbf{v} \in \mathcal{V}$, all the minimal constraints sets forming the optimization problem whose unique optimal solution is \mathbf{v} ; denote this family by $\mathcal{F}(\mathbf{v}) = \{F : F \subset \mathcal{H}(\mathbf{v}), \mathbf{v} = \boldsymbol{\mu}_F, \mathbf{v} \neq \boldsymbol{\mu}_{F \setminus \{j\}} \text{ for all } j \in F\}$. Note that $|F| \leq d$ for each $F \in \mathcal{F}(\mathbf{v})$ by Lemma 1 and hence the cardinality of $\bigcup_{\mathbf{v} \in \mathcal{V}} \mathcal{F}(\mathbf{v})$ has the same upper bound as the one in Lemma 2. Because we search \mathcal{V} by probing all index sets of cardinality less than or equal to d , we get $\mathcal{F}(\mathbf{v})$'s as by-products of the search. To simplify notations, we abuse the symbol \mathcal{V} to denote the collection of pairs (\mathbf{v}, F) for each \mathbf{v} and each $F \in \mathcal{F}(\mathbf{v})$.

To identify $\{\boldsymbol{\mu}_{\mathcal{J}} : \mathcal{J} \in \mathcal{S}_q\}$ from \mathcal{V} according to our scheme, we have to decide whether there is a $\mathcal{J} \in \mathcal{S}_q$ such that $\mathbf{v} = \boldsymbol{\mu}_{\mathcal{J}}$ for each $\mathbf{v} \in \mathcal{V}$. For this decision, we can use some information on \mathbf{v} , $\mathcal{H}(\mathbf{v})$ and $\mathcal{F}(\mathbf{v})$, which can be collected from the computation of \mathcal{V} with no additional cost. We formulate this problem as a *minimal cover* problem (MCP). Then we transform MCP into a knapsack problem. To simplify notations, we define $C_J := \sum_{j \in J} C_j$ for any index set J .

MCP: An index set N is given. $\{C_i\}_{i \in N}$ with $C_i > 0$ and a subset $F \subset N$ ($F \neq \emptyset$) are given. For a given positive number b , is there a subset $J \subset N \setminus F$ such that

$$C_{J \cup F} \geq b \quad \text{and} \quad C_{J \cup F \setminus \{k\}} < b \text{ for all } k \in J \cup F ?$$

Then we have the following lemma:

Lemma 3. *The answer to MCP is YES if and only if there exists a $J \subset N \setminus F$ such that*

$$\text{i) } C_{J \cup F} \geq b, \text{ ii) } C_{J \cup F \setminus \{k\}} < b \text{ for all } k \in J, \text{ and iii) } C_{J \cup F} - \min_{i \in F} C_i < b.$$

Proof. If we notice the relation $C_{J \cup F \setminus \{k\}} < b$ for all $k \in F \Leftrightarrow C_{J \cup F} - \min_{i \in F} C_i < b$, then the proof is complete. \square

Set $b' := b - C_F$. Using Lemma 3, we can rewrite the MCP as

MCP': $\{C_i\}_{i \in N}$ with $C_i > 0$ and a subset $F \subset N$ are given. For a given positive number b , is there a subset $J \subset N \setminus F$ such that

$$\text{i) } C_J \geq b', \text{ ii) } C_{J \setminus \{k\}} < b' \text{ for all } k \in J, \text{ and iii) } C_J < b' + \min_{i \in F} C_i ?$$

Consider the following 0-1 knapsack problem (KP):

$$f^* = \min \left\{ \sum_{j \in N \setminus F} C_j x_j : \sum_{j \in N \setminus F} C_j x_j \geq b', x_j \in \{0, 1\} \text{ for all } j \in N \setminus F \right\}.$$

Any set $G \subset N \setminus F$ corresponding to an optimal solution of (KP) satisfies condition i) of MCP' from the feasibility. If $C_{G \setminus \{k\}} \geq b'$ for some $k \in G$, then $G \setminus \{k\}$ is another feasible set with strictly less optimal value and this contradicts to the optimality of G . Hence G satisfies ii) of MCP'. Therefore, we conclude that $f^* < b' + \min_{i \in F} C_i$ if and only if the answer to MCP is YES. Now set $N = \mathcal{H}(\mathbf{v})$ and take an F from $\mathcal{F}(\mathbf{v})$. Then by setting $b = qC$, MCP solves whether there is a \mathcal{J} such that $F \subset \mathcal{J} \subset \mathcal{H}(\mathbf{v})$, $\mathcal{J} \in \mathcal{S}_q$, and $\mathbf{v} = \boldsymbol{\mu}_{\mathcal{J}}$. Hence by checking this question for all $F \in \mathcal{F}(\mathbf{v})$, we can decide whether $\mathbf{v} \in \{\boldsymbol{\mu}_{\mathcal{J}} : \mathcal{J} \in \mathcal{S}_q\}$. Note that, with $\min_{i \in F} C_i = 1$, MCP' is equivalent to knapsack feasibility problem and hence MCP is NP-complete.

By transforming MCP' into the maximization form using the minimal index set notations results in the following SSP:

$$f^* = \max \left\{ \sum_{j \in N \setminus F} C_j x_j : \sum_{j \in N \setminus F} C_j x_j \leq C_N - qC, x_j \in \{0, 1\} \text{ for all } j \in N \setminus F \right\}.$$

The procedure identifying $\{\boldsymbol{\mu}_{\mathcal{J}} : \mathcal{J} \in \mathcal{S}_q\}$ is described as the following:

-
- 1: Identify \mathcal{V} by solving the norm minimization problems (2) associated with all possible combinations of type indices, $\mathcal{J} \subset \{1, \dots, t\}$, $|\mathcal{J}| \leq d$.
 - 2: Given q , solve SSP associated with each $N = \mathcal{H}(\mathbf{v})$ and $F \in \mathcal{F}(\mathbf{v})$ for all $\mathbf{v} \in \mathcal{V}$. If $f^* > C_N - qC - \min_{i \in F} C_i$, then include \mathbf{v} among the shifting mean vectors.
-

We assume that all C_j 's are positive integers. This is a necessary assumption for knapsack problems. SSP has a special structure and is called a *subset sum* problem which is NP-complete. However, knapsack problems arising in practice are solved very fast. (See, e.g., Chapter 4 of Kellerer, Pferschy, and Pisinger [7].) For numerical experiment, we measured the time spent to solve 10^6 subset sum problems using a code `subsum.c` available at <http://www.diku.dk/~pisinger>. Each instance consists of 100 randomly generated weights (i.e. $|N \setminus F| = 100$ in SSP) and the weights have their ranges $[1, 10^4]$ (i.e., $1 \leq C_j \leq 10^4$). 21.88 seconds were spent to solve all these 10^6 problems. (All experiments in this paper were executed using a notebook with a CPU of 1.7GHz Intel Pentium M and a 512MB RAM.) This number of problems, 10^6 , is roughly the upper bound of the cardinality of \mathcal{V} for a factor model having 100 types ($= |N|$) and three factors. In solving a subset sum problem, the range of weights are crucial for the running time of algorithm. The above input ranges imply that the potential loss amount of each obligor will take its value among the multiples up to 10^4 of some base amount.

Table 1 shows the average cardinalities of $\{\mu_{\mathcal{J}} : \mathcal{J} \in \mathcal{S}_q\}$ and \mathcal{V} for 30 randomly generated 20- or 25-type instances with factor dimension 4 or 5. Note that the values of the upper bound on $|\mathcal{V}|$ in Lemma 2 are 6195, 21699, 15275, and 68405, respectively. However we just need to keep a smaller size (at most 2000 on average) of \mathcal{V} to get $\{\mu_{\mathcal{J}} : \mathcal{J} \in \mathcal{S}_q\}$. The computing time of \mathcal{V} takes about 28, 100, 65, and 300 seconds for each instance, respectively if we use the MATLAB function `quadprog` for the norm minimization (2). (By a specialized algorithm in Section 3.3, the time can be reduced to 0.3, 1, 1, and 6 seconds for each instance, respectively.) And the total times in solving 9 subset sum problems to find $\{\mu_{\mathcal{J}} : \mathcal{J} \in \mathcal{S}_q\}$'s for $q = 0.1, 0.2, \dots, 0.9$ from \mathcal{V} are at most 0.2, 0.5, 0.4, and 1.5 seconds, respectively. Furthermore, the cardinalities of $\{\mu_{\mathcal{J}} : \mathcal{J} \in \mathcal{S}_q\}$ are much smaller than the theoretical upper bound. These observations imply that we can implement the IS efficiently.

Table 1. The average number of minimum norm points in \mathbb{R}^d . n_q denotes the average of $|\{\mu_{\mathcal{J}} : \mathcal{J} \in \mathcal{S}_q\}|$. (This table is taken from [3]).

Types	d	Bound	$ \mathcal{V} $	$n_{0.1}$	$n_{0.2}$	$n_{0.3}$	$n_{0.4}$	$n_{0.5}$	$n_{0.6}$	$n_{0.7}$	$n_{0.8}$	$n_{0.9}$
20	4	6195	574.6	16.9	36.1	48.5	52.2	44.5	29.6	14.3	3.9	0.2
20	5	21699	932.2	25.0	57.0	78.8	84.4	69.0	44.2	19.5	4.9	0.4
25	4	15275	1224.9	33.5	65.7	90.5	91.7	74.6	44.1	16.0	2.4	0.2
25	5	68405	2036.5	39.7	96.3	138.4	157.1	137.7	79.8	28.2	3.1	0.0

3.3 Quadratic Optimizations

To find \mathcal{V} , we need to solve (2). We can apply general quadratic programming (QP) algorithms to these problems. However, we can exploit the hierarchy of QP problems further: we characterize \mathcal{V} by solving a QP for each $\mathcal{J} \subset \{1, \dots, t\}$, $|\mathcal{J}| \leq d$. This strategy of the search allows us to do it by

solving $\nu_{\mathcal{J}} = \operatorname{argmin}\{\|\mathbf{z}\| : \mathbf{a}_j^\top \mathbf{z} = d_j \text{ for all } j \in \mathcal{J}\}$ instead of the original QP contrained by inequalities. This equality constrained problem can be solved by simple Gaussian eliminations. Because of the change of constraints, we have $\|\nu_{\mathcal{J}}\| \geq \|\mu_{\mathcal{J}}\|$ instead of equality. So we have to detect the case $\|\nu_{\mathcal{J}}\| > \|\mu_{\mathcal{J}}\|$. For this, we adopt the following procedure:

```

Set  $L = \emptyset$ 
for  $i = 1$  to  $d$ 
  for all  $\mathcal{J} \subset \{1, \dots, t\}$  of  $|\mathcal{J}| = i$ 
    • find  $\nu_{\mathcal{J}}$ 
    • check the existence of  $\mathcal{J}' \in L$  so that  $\mathcal{J}' \subset \mathcal{J}$  and  $\nu_{\mathcal{J}'} \leq \nu_{\mathcal{J}}$ 
    • if no such  $\mathcal{J}'$  then add  $\mathcal{J}$  to  $L$ .
  end
end

```

Note that there always exists a $\mathcal{J}' \subset \mathcal{J}$ such that $\nu_{\mathcal{J}'} = \mu_{\mathcal{J}'}$ if $\|\nu_{\mathcal{J}}\| > \|\mu_{\mathcal{J}}\|$. Furthermore, $\nu_{\mathcal{J}} = \mu_{\mathcal{J}}$. Since the enumeration is done in increasing order of $|\mathcal{J}|$, $\nu_{\mathcal{J}}$ exists in the list L (because $|\mathcal{J}'| < |\mathcal{J}|$). Hence the \mathcal{J} is discarded before we solve (2) for \mathcal{J} . By this implementation, we can reduce substantial amount of time spent to identify \mathcal{V} .

4 Concluding Remarks

We considered an optimization problem arising in the simulation of portfolio credit risk. Our re-formulation has the same worst case computational complexity as the original problem, but it allows tractability in practice. The shifting of sampling distribution based on these points enhances the efficiency of simulation quite impressively.

Acknowledgments

The first author thanks Paul Glasserman and the late Perwez Shahabuddin, the coauthors of [3], on which this proceeding is based.

References

1. A. Avranitis and J. Gregory. *Credit: The Complete Guide to Pricing, Hedging and Risk Management*. Risk Books, London, 2001.
2. P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer-Verlag, New York, 2004.
3. P. Glasserman, W. Kang, and P. Shahabuddin. Fast simulation of multifactor portfolio credit risk. Technical report, Graduate School of Business and IEOR Department, Columbia University, February 2005.
4. P. Glasserman and J. Li. Importance sampling for portfolio credit risk. *Management Science*, 2005.

5. G. Gupton, C. Finger, and M. Bhatia. *CreditMetrics Technical Document*. J.P. Morgan & Co., New York, NY, 1997.
6. M. Kalkbrener, H. Lotter, and L. Overbeck. Sensible and efficient capital allocation for credit portfolios. *RISK*, January:S19–S24, 2004.
7. H. Kellerer, U. Pferschy, and D. Pisinger. *Knapsack Problems*. Springer-Verlag, Berlin · Heidelberg, Germany, 2004.
8. S. Merino and M. A. Nyfeler. Applying importance sampling for estimating coherent credit risk contributions. *Quantitative Finance*, 4:199–207, 2004.
9. W. Morokoff. An importance sampling method for portfolios of credit risky assets. In R. Ingalls, M. Rossetti, J. Smith, and B. Peters, editors, *Proceedings of the 2004 Winter Simulation Conference*, pages 1668–1676, 2004.
10. J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer-Verlag, New York, 1999.

Two-Server Network Disconnection Problem^{*}

Byung-Cheon Choi and Sung-Pil Hong^{**}

Department of Industrial Engineering, College of Engineering,
Seoul National University, San 56-1, Shillim-9-Dong, Seoul, 151-742, Korea
sphong@snu.ac.kr

Abstract. Consider a set of users and servers connected by a network. Each server provides a unique service which is of certain benefit to each user. Now comes an attacker, who wishes to destroy a set of edges of the network in the fashion that maximizes his net gain, namely, the total disconnected benefit of users minus the total edge-destruction cost. We first discuss that the problem is polynomially solvable in the single-server case. In the multiple-server case, we will show, the problem is, however, *NP*-hard. In particular, when there are only two servers, the network disconnection problem becomes intractable. Then a $\frac{3}{2}$ -approximation algorithm is developed for the two-server case.

1 Introduction

Consider a network of servers and their users in which each server provides a unique service to the users. (Each server also can be the user of a service other than her own.) Each user takes a benefit through the connection provided by the network to each server. Now comes an attacker who wishes to destroy a set of edges in the manner that optimizes his own objective, although defined to various circumstances, that explicitly accounts for the disconnected benefits of users resulted from destruction.

Such a model, to the author's best knowledge, was proposed first by Martel et al. (8). They considered the single-server problem in which the total disconnected benefit is maximized under an edge-destruction budget constraint. The problem, as they showed, is *NP*-hard. They also proposed an exact method enumerating maximum disconnected node sets among minimum cost cuts separating node pairs using cut submodularity.

In this paper, we consider the problem of maximizing net gain, namely, the total disconnected benefit of users minus the edge-destroying cost. The problem is polynomially solvable when there is a single server. It is, however, *NP*-hard in general. We will provide the proofs. In particular, if there are two servers, the problem becomes intractable. Also, we will present a $\frac{3}{2}$ -approximation algorithm for the two-server case.

^{*} The research was partially supported by KOSEF research fund R01-2005-000-10271-0.

^{**} Corresponding author.

In Section 2, we review the previous models on separating nodes of a graph. Section 3 provides a mathematical model of the problem. It also discusses the polynomiality of the single-server case and *NP*-hardness for the *k*-server case with $k \geq 2$. A $\frac{3}{2}$ -approximation algorithm is presented in Section 4. Finally, we summarize the results and point out an open problem in Section 5.

2 Node Separation Problems: A Literature Review

In this section we review combinatorial optimization models on the separation of nodes of an undirected graph. There are various such models. (See, e.g. (1; 5).) Among them, the following three models probably have been studied most intensively.

Problem 1. *k*-cut problem: Given an undirected $G = (N, E)$ with nonnegative edge weights, find a minimum weight set of edges E' such that the removal of E' from E separates graph into exactly *k* nonempty components.

This problem is *NP*-hard for general $k \geq 3$ and approximable within factor $2 - \frac{2}{k}$ within the optimum. Goldschmidt and Hochbaum (7) showed that the problem is polynomially solvable for fixed *k*.

Problem 2. *k*-terminal cut problem, Multiway cut problem: Given an undirected $G = (N, E)$ with nonnegative edge weights and a set of *k* specified nodes, or *terminals*, find a minimum weight set of edges E' such that the removal of E' from E disconnects each terminal from all the others.

k-terminal cut problem is *Max-SNP*-hard even for $k = 3$ (4). Therefore, there is some $\epsilon > 0$ such that $(1 + \epsilon)$ -approximation is *NP*-hard. Naor and Zosin(9) considered the directed version of *k* multi-terminal cut problem, and presented two 2-approximation algorithms. The current best approximation guarantee is $\frac{3}{2} - \frac{1}{k}$ by Călinescu et al. (3).

	<i>k</i> -cut problem	<i>k</i> -terminal cut problem	multicut problem
$k = 2$	P^a	P^b	$P(12)$
$k \geq 3$	$P(2; 7)$	Max- <i>SNP</i> -hard ^{c,d} (4)	Max- <i>SNP</i> -hard ^e
arbitrary	<i>NP</i> -hard ^f (7)	Max- <i>SNP</i> -hard ^{g,h} (4; 3)	Max- <i>SNP</i> -hard ⁱ

^a *P* means “polynomially solvable”

^b *Max-SNP*-hard if there is a constraint on the size of the separated component

^c polynomially solvable if the graph is planar

^d *NP*-hard even if all edge weights are equal to 1

^e *NP*-hard by the reduction from 3-terminal cut problem

^f There is a $(2 - \frac{2}{k})$ -approximation algorithm

^g There is a $(\frac{3}{2} - \frac{1}{k})$ -approximation algorithm

^h *NP*-hard even if the graph is planar and all edge weights are equal to 1

ⁱ approximable within $O(\log k)$.

Problem 3. Multicut problem: Given an undirected $G = (N, E)$ with nonnegative edge weights and k pairs of nodes, find a minimum weight set of edges E' such that the removal of E' from E separates all pairs of nodes.

This problem is a generalization of the k -terminal cut problem. Hence, it also becomes Max-SNP-hard when $k = 3$. Currently, this problem is approximable within the factor of $O(\log k)$ (6).

These works can be summarized as above text table.

In the NP-hardness proof of the k -server network disconnection problem, we reduce the $(k+1)$ -terminal problem to the k -server network disconnection problem.

3 The k -Server Network Disconnection Problem

3.1 Problem Formulation

Given an undirected graph $G = (N, E)$ with $N = \{1, 2, \dots, n\}$, the server set $S = \{s_1, \dots, s_k\} \subseteq N$, the costs $c_{ij} \geq 0$, $(i, j) \in E$, the nonnegative vectors $d_i = (d_i^1, \dots, d_i^k)^T$, $i \in N$, find a set $F \subseteq E$ that maximizes the total disconnected benefits of nodes minus the edge-destruction cost, $\sum_{(i,j) \in F} c_{ij}$.

Let N_l be the set of nodes remaining connected from server l after the destruction of F . Then, it is easy to see that for any $i \neq j$, the two sets, N_i and N_j are either identical or mutually exclusive: $N_i = N_j$ or $N_i \cap N_j = \emptyset$. Hence, if we denote by N_0 the nodes disconnected from all the servers, then the set $\mathcal{N} = \{N_0, N_1, \dots, N_k\}$ is a partition of N . If \mathcal{N} has p distinct sets, we will call \mathcal{N} a p -partition. Our problem can be restated as follows:

Problem 1. k -server network disconnection problem: Find a partition $\mathcal{N} = \{N_0, N_1, \dots, N_k\}$ with $s_j \in N_j$, $j = 1, 2, \dots, n$ that maximizes

$$z(\mathcal{N}) = \sum_{l: N_l \in \mathcal{N}} \sum_{i \in N_l} \sum_{j \notin N_l} d_i^j - \sum_{\substack{(i,j) \in E \\ \text{such that } N_p = N_q}} c_{ij}. \tag{1}$$

3.2 Polynomiality of the Single-Server Case

We need to find $N_1 \subseteq N$ with $s_1 \in N_1$ so that $\mathcal{N} = \{N \setminus N_1, N_1\}$ maximizes (1). Define a binary variable x as follows:

$$x_j = \begin{cases} 1, & \text{if } j \in N_1, \\ 0, & \text{otherwise.} \end{cases}$$

For notational convenience we assume $s_1 = 1$, $c_{ij} = c_{ji}$ for all $i, j \in N$ and $c_{ij} = 0$ for $(i, j) \notin E$. Then we can formulate the single-server problem as a 0-1 quadratic program as follows.

$$\begin{aligned} \max z(x) &= \sum_{j=1}^n d_j^1(1 - x_j) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n c_{ij}(x_i - x_j)^2 \\ \text{sub. to } &x_1 = 1, x_j \in \{0, 1\}, j \in N \setminus \{1\}. \end{aligned} \tag{2}$$

Using $x_j = x_j^2$, it is easy to rewrite the objective of (2) as a quadratic function of $\hat{x} = (x_2, x_3, \dots, x_n)$: $z(\hat{x}) = \hat{x}^T Q \hat{x} + \sum_{j=2}^n (d_j^1 - c_{1j})$, where $Q = (q_{ij})_{\substack{i=2, \dots, n \\ j=2, \dots, n}}$ is given as follows:

$$q_{ij} = \begin{cases} -d_i^1 + c_{i1} - \sum_{k=2}^n c_{ik}, & \text{if } i = j, \\ c_{ij}, & \text{otherwise.} \end{cases} \tag{3}$$

Lemma 1. *An unconstrained 0-1 quadratic maximization with nonnegative off-diagonal elements is polynomially solvable.*

Proof. See Picard and Ratliff (10). □

Theorem 1. *The single-server network disconnection problem is polynomially solvable.*

Proof. Since $c_{ij} \geq 0$, $(i, j) \in E$, the polynomiality of the quadratic program formulation (2), follows from (3) and Lemma 1. □

3.3 NP-Hardness of the k-Server Case

Unlike the single-server case, our problem is NP-hard in general. When $k = 2$, in particular, it becomes NP-hard. To see this, consider the decision version of the 3-terminal cut problem, Problem 2.

Problem 2. Decision version of 3-terminal cut problem(3DMC) Given a terminal set $T = \{t_1, t_2, t_3\}$, a weight $w_{ij} \geq 0$, $(i, j) \in E$ and a constant W , is there a partition $\mathcal{N} = (N_1, N_2, N_3)$ such that $t_j \in N_j$, $j = 1, 2, 3$, and

$$w(\mathcal{N}) = \sum_{\substack{(i,j) \in E \\ \text{such that } N_p = N_q}} w_{ij} \leq W?$$

Theorem 2. *The 2-server network disconnection problem is NP-hard.*

Proof. Given any instance of 3DMC, we can construct an instance of the 2-server problem as follows : On the same graph $G = (N, E)$, designate the first two terminals of 3DMC as the two server nodes, $s_1 = t_1$ and $s_2 = t_2$ while t_3 is a non-server node in the 2-server instance. Now, define the benefit vector for the node of the 2-server instance: For a constant $M > W$, set $d_{s_1} = (d_{s_1}^1, d_{s_1}^2) = (0, M)$, $d_{s_2} = (M, 0)$, and $d_{t_3} = (M, M)$. The other nodes are assigned to the benefit vector, $(0, 0)$. Also set $Z = 4M - W$. Finally, the edge cost is defined as $c_{ij} = w_{ij}$, $(i, j) \in E$. We claim that the answer to 3DMC is ‘yes’ if and only if the 2-server instance has a partition whose value is no less than Z .

Suppose there is a partition $\hat{\mathcal{N}} = \{\hat{N}_1, \hat{N}_2, \hat{N}_3\}$ of 3DMC satisfying $w(\hat{\mathcal{N}}) \leq W$. Then, it is easy to see that $\mathcal{N} = \{\hat{N}_3, \hat{N}_1, \hat{N}_2\}$ is a solution of the 2-server instance: $s_j \in \hat{N}_j$, $j = 1, 2$, and

$$z(\hat{\mathcal{N}}) = \sum_{l=1}^3 \sum_{i \in \hat{N}_l} \sum_{j \notin \hat{N}_l} d_i^j - \sum_{\substack{(i,j) \in E \\ \text{such that } \hat{N}_p = \hat{N}_q}} w_{ij} \geq 4M - W = Z.$$

On the other hand, assume $\tilde{\mathcal{N}} = \{\tilde{N}_0, \tilde{N}_1, \tilde{N}_2\}$ is a solution of the 2-server instance such that $z(\tilde{\mathcal{N}}) \geq Z$. To show that $\tilde{\mathcal{N}}$ is a 3-terminal cut, it suffices to

see that $\tilde{N}_j, j = 0, 1, 2$ are all pair-wise distinct. But, if any two of the sets are identical, then $z(\tilde{x}) \leq 3M$, a contradiction to the assumption. Furthermore, this also implies

$$z(\tilde{\mathcal{N}}) = 4M - w(\tilde{x}) \geq Z = 4M - W.$$

Since $4M - w(\tilde{\mathcal{N}}) \geq Z = 4M - W$, we get $w(\tilde{\mathcal{N}}) \leq W$. □

It is not hard to see that the proof can be extended to show that the $(k + 1)$ -terminal cut problem is a special case of the k -server network disconnection problem.

4 A $\frac{3}{2}$ -Approximation of the 2-Server Case

Let $\mathcal{N} = \{N_0, N_1, N_2\}$ be a solution of the 2-server case such that $s_1 \in N_1$ and $s_2 \in N_2$. Define

$$\begin{aligned} E_0 &= \{(i, j) \in E \mid i \in N_1 \text{ and } j \in N_2\} \\ E_1 &= \{(i, j) \in E \mid i \in N_1 \text{ and } j \in N_0\}, \text{ and} \\ E_2 &= \{(i, j) \in E \mid i \in N_2 \text{ and } j \in N_0\}. \end{aligned}$$

Recall that it may be either $N_1 = N_2$ or $N_0 = \emptyset$ and thus an optimal solution

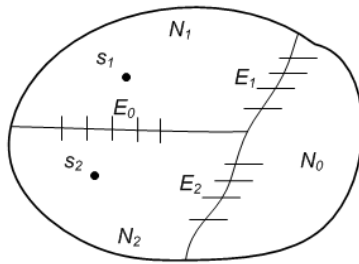


Fig. 1. 2-server solution

may be 1-, 2-, or 3-partition. If it is trivially a 1-partition, then the optimal objective value is 0. The idea is to approximate the optimum with an optimal 2-partition which is, as we will show, polynomially computable.

Algorithm H : Find an optimal 2-partition as an approximation of the optimum.

Lemma 1. *An optimal 2-partition can be computed in polynomial time.*

Proof. There are two cases of an optimal 2-partition: $N_1 = N_2$ or $N_1 \neq N_2$.

Case 1: N_1 and N_2 are distinct

Then we can formulate the 2-server problem as a quadratic program similarly to the single-server case. Define a binary variable x as follows:

$$x_j = \begin{cases} 1, & \text{if } j \in N_1, \\ 0, & \text{if } j \in N_2. \end{cases}$$

As before, we adopt the notation, $s_1 = 1, s_2 = 2, c_{ij} = c_{ji}$ for all $i, j \in N$ and $c_{ij} = 0, (i, j) \notin E$. Then, it is easy to see that the 2-server case is equivalent to

$$\begin{aligned} \max \quad & z(x) = \sum_{j=1}^n (d_j^2 x_j + d_j^1 (1 - x_j)) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n c_{ij} (x_i - x_j)^2 \quad (4) \\ \text{sub. to} \quad & x_1 = 1, x_2 = 0, \\ & x_j \in \{0, 1\}, j \in N \setminus \{1, 2\}. \end{aligned}$$

Then we can rewrite (4) in terms of $\hat{x} = (x_3, x_4, \dots, x_n)^T: z(\hat{x}) = \hat{x}^T Q \hat{x} + \sum_{j=1}^n d_j^1 + d_1^2 + d_2^2 - 2c_{12}$, where $Q = (q_{ij})_{\substack{i=3, \dots, n \\ j=3, \dots, n}}$ is given as follows:

$$q_{ij} = \begin{cases} d_i^1 - d_i^2 + c_{i1} - c_{i2} - \sum_{k=1}^n c_{ik}, & \text{if } i = j, \\ c_{ij}, & \text{otherwise.} \end{cases} \quad (5)$$

Since $c_{ij}, (i, j) \in E$ are nonnegative, $z(\hat{x})$ can be solved in polynomial time from Lemma 1.

Case 2: N_1 and N_2 are identical

In this case, the problem is essentially a single-server problem. Merge s_1 and s_2 into a single node and replace benefit vector (d_i^1, d_i^2) by a single benefit $d_i^1 + d_i^2$. Then, it is easy to see that the 2-server case is equivalent to the single-server case defined on the modified network, implying that the 2-server case can be solved in polynomial time. Case 1 and 2 complete proof. \square

Theorem 1. *An optimal 2-partition is factor $\frac{3}{2}$ within the optimum.*

Proof. Let $\mathcal{N}^* = \{N_0^*, N_1^*, N_2^*\}$ and $\mathcal{N}^H = \{N_0^H, N_1^H, N_2^H\}$ be an optimal solution and the solution of *Algorithm H*, respectively.

$$\begin{aligned} z(\mathcal{N}^*) &= \sum_{l=0}^2 \sum_{i \in N_l} \left(\sum_{j=1}^2 d_i^j - d_i^l \right) - \sum_{(i,j) \in E_0 \cup E_1 \cup E_2} c_{ij} \\ &= \frac{1}{2} \left(\left(\sum_{i \in N_0} d_i^2 + \sum_{i \in N_2} d_i^1 - \sum_{(i,j) \in E_0, E_2} c_{ij} \right) + \left(\sum_{i \in N_0, N_2} d_i^1 + \sum_{i \in N_1} d_i^2 - \sum_{(i,j) \in E_0, E_1} c_{ij} \right) \right. \\ &\quad \left. + \left(\sum_{i \in N_0} (d_i^1 + d_i^2) - \sum_{(i,j) \in E_1 \cup E_2} c_{ij} \right) \right) \\ &= \frac{1}{2} (z(N_0^*, N_1^* \cup N_2^*) + z(N_1^*, N_0^* \cup N_2^*) + z(N_2^*, N_0^* \cup N_1^*)). \end{aligned}$$

Since $z(\mathcal{N}^H) \geq \max\{z(N_0^*; N_1^* \cup N_2^*), z(N_1^*; N_0^* \cup N_2^*), z(N_2^*; N_0^* \cup N_1^*)\}$,

$$z(\mathcal{N}^*) \leq \frac{3}{2}z(\mathcal{N}^H).$$

This completes the proof. \square

Theorem 2. *Algorithm H is a $\frac{3}{2}$ -approximation algorithm for the 2-server case.*

Proof. From Lemma 1 and Theorem 1. \square

5 Summary and Further Research

We consider the k -server network disconnection problem. We show that the problem can be solved in polynomial time for the single-server case while the problem is NP-hard in general. The problem is NP-hard even for the two-server case. Also, we propose a $\frac{3}{2}$ -approximation algorithm for two-server case.

The approximability of the k -server network disconnection problem for general $k \geq 0$ remains open. The quadratic program based approximation algorithm for the 2-server case does not seem to straightforwardly extend to the general case.

Acknowledgment

The authors thank Prof. Young-Soo Myung for the fruitful discussion which leads to a tighter analysis of the algorithm.

References

- [1] G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi. *Complexity and Approximation*, Springer-Verlag, Berlin, 1999.
- [2] M. Buriel, and O. Goldschmidt. *A new and improved algorithm for the 3-cut problem*. Operations Research Letters 21(1997) pp. 225-227.
- [3] G. Călinescu, H. Karloff, and Y. Rabani. *An improved approximation algorithm for multiway cut*, Proc. 30th ACM Symposium on Theory of Computing, ACM, (1998) pp. 48-52.
- [4] E. Dahlhaus, D.S. Johnson, C.H. Papadimitriou, P.D. Seymour, and M. Yannakakis. *The complexity of multiterminal cuts*, SIAM Journal on Computing 23(1994) pp. 864-894.
- [5] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*, W.H. Freeman, New York, 1979.
- [6] N. Garg, V.V. Vazirani, and M. Yannakakis. *Approximating max-flow min-(multi)cut theorems and their applications*, SIAM Journal on Computing 25(1996) pp. 235-251.
- [7] O. Goldschmidt, and D.S. Hochbaum. *A polynomial time algorithm for the k-cut problem for k fixed*, Mathematics of Operations Research 19(1994) pp. 24-37.

- [8] C. Martel, G. Nuckolls, and D. Sniegowski. *Computing the disconnectivity of a graph*, manuscript, UC Davis (2001).
- [9] J. Maor, and L. Zosin. *A 2-approximation algorithm for the directed multiway cut problem*, SIAM Journal on Computing 31(2001) pp. 477-482.
- [10] J.C. Picard, and H.D. Ratliff. *Minimum cuts and related problems*, Networks 5(1974) pp. 357-370.
- [11] H. Saran, and V.V. Vazirani. *Finding k -cuts within twice the optimal*, Proc. 32nd Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, (1991) pp. 743-751.
- [12] M. Yannakakis, P.C. Kanellakis, S.C. Cosmadakis, and C.H. Papadimitriou. *Cutting and partitioning a graph after a fixed pattern*, in *Automata, Language and Programming*, Lecture Notes and Computer Science 154(1983) pp. 712-722.

One-Sided Monge TSP Is NP-Hard

Vladimir Deineko¹ and Alexander Tiskin²

¹ Warwick Business School, The University of Warwick, Coventry CV47AL, UK

V.Deineko@warwick.ac.uk

² Dept. of Computer Science, The University of Warwick, Coventry CV47AL, UK

tiskin@dcs.warwick.ac.uk

Abstract. The Travelling Salesman Problem (TSP) is a classical NP-hard optimisation problem. There exist, however, special cases of the TSP that can be solved in polynomial time. Many of the well-known TSP special cases have been characterized by imposing special *four-point conditions* on the underlying distance matrix. Probably the most famous of these special cases is the TSP on a *Monge matrix*, which is known to be polynomially solvable (as are some other generally NP-hard problems restricted to this class of matrices). By relaxing the four-point conditions corresponding to Monge matrices in different ways, one can define other interesting special cases of the TSP, some of which turn out to be polynomially solvable, and some NP-hard. However, the complexity status of one such relaxation, which we call *one-sided Monge TSP* (also known as the TSP on a *relaxed Supnick matrix*), has remained unresolved. In this note, we show that this version of the TSP problem is NP-hard. This completes the full classification of all possible four-point conditions for symmetric TSP.

1 Introduction

The travelling salesman problem (TSP) is a well-known problem of combinatorial optimisation. In the symmetric TSP, given a symmetric $n \times n$ distance matrix $C = (c_{ij})$, one looks for a cyclic permutation τ of the set $\{1, 2, \dots, n\}$ that minimises the function $c(\tau) = \sum_{i=1}^n c_{i, \tau(i)}$. The value $c(\tau)$ is called the *length* of the permutation τ . We will in the following refer to the items in τ as *points*.

The TSP is an NP-hard problem [10]. There exist, however, special cases of the TSP that can be solved in polynomial time. For a survey of efficiently solvable cases of the TSP, see [3, 11, 14]. Many of the well-known TSP special cases result from imposing special conditions on the underlying distance matrix. Probably the most famous of these special cases is the TSP on a Monge matrix. For a number of well-known NP-hard problems, including TSP, restriction to Monge matrices reduces the complexity to polynomial (see survey [5]).

We give below two equivalent definitions of a Monge matrix. In what follows, we will always assume that the matrices under considerations are symmetric.

Definition 1. An $n \times n$ matrix $C = (c_{ij})$ is a Monge matrix, if it satisfies the Monge conditions:

$$c_{ij} + c_{i'j'} \leq c_{ij'} + c_{i'j} \quad 1 \leq i < i' \leq n \quad 1 \leq j < j' \leq n \quad (1)$$

Definition 2. An $n \times n$ matrix $C = (c_{ij})$ is a Monge matrix, if

$$c_{ij} + c_{i+1,j+1} \leq c_{i,j+1} + c_{i+1,j} \quad 1 \leq i \leq n-1 \quad 1 \leq j \leq n-1$$

It can be easily seen that the above two definitions define the same set of matrices. The difference arises when we begin to generalize the problem by relaxing the conditions imposed on the matrix. For example, the diagonal entries are not involved in the calculation of the TSP objective function, so one can define c_{ii} ($i = 1, \dots, n$) arbitrarily without affecting the solution of the TSP. If we relax Definition 1 by excluding the inequalities containing the diagonal entries, then the structural properties of the matrix remain essentially unchanged. In fact, given a matrix (c_{ij}) with indeterminate diagonal elements, which satisfies inequalities (1) for $i \neq j, i \neq j', i' \neq j, i' \neq j'$, it is always possible to define diagonal elements c_{ii} so that the resulting matrix is a Monge matrix (see Proposition 2.13 in [3]). This relaxation of symmetric Monge matrices is known as *Supnick matrices*. Alternatively, Supnick matrices are defined as matrices satisfying conditions

$$c_{ij} + c_{j+1,l} \leq c_{i,j+1} + c_{jl} \leq c_{il} + c_{j,j+1} \quad 1 \leq i < j < j+1 < l \leq n$$

Supnick [21] has shown that an optimal TSP tour on such matrices is given by $\sigma_{Smin} = \langle 1, 3, 5, 7, 9, \dots, 8, 6, 4, 2, 1 \rangle$.

The well-known classes of Demidenko [9] and Van der Veen [23] matrices, as well as a class of matrices investigated in [4], can also be viewed as relaxations of Definition 1. A symmetric $n \times n$ matrix $C = (c_{ij})$ is a *Demidenko matrix*, if

$$c_{ij} + c_{i'j} \leq c_{ij} + c_{i'j} \quad 1 \leq j < i < i' < j' \leq n$$

and a *Van der Veen matrix*, if

$$c_{ij} + c_{i'j} \leq c_{ij} + c_{i'j} \quad 1 \leq i < j < i' < j' \leq n$$

The TSP on these classes of matrices is polynomially solvable: an optimal tour can be found among the so-called pyramidal tours in $O(n^2)$ time (see e.g. [3]).

Now, instead of Definition 1, we consider Definition 2, and relax it by excluding inequalities involving the diagonal elements.

Definition 3. An $n \times n$ matrix $C = (c_{ij})$ is a relaxed Supnick matrix, if

$$c_{ij} + c_{i+1,j+1} \leq c_{i,j+1} + c_{i+1,j} \quad 1 \leq i < j-1 \leq n-2$$

In contrast with the Supnick relaxation of Definition 1, the relaxation from Definition 2 to Definition 3 may have a very significant effect on the matrix structure. While the Supnick conditions constrain all pairs of matrix elements excluding the main diagonal, the relaxed Supnick conditions only constrain pairs of matrix elements that are on the same side of the main diagonal (i.e. are either both above, or, by symmetry, both below the main diagonal). In calling such matrices “relaxed Supnick”, we follow the terminology of [6]; otherwise, they could naturally be called *one-sided Monge matrices*.

Table 1. Classification of four-point conditions (adapted from [6])

	$\mathcal{A} \leq \mathcal{B}$	$\mathcal{A} \geq \mathcal{B}$	$\mathcal{A} \leq \mathcal{C}$	$\mathcal{A} \geq \mathcal{C}$	$\mathcal{B} \leq \mathcal{C}$	$\mathcal{B} \geq \mathcal{C}$
$\mathcal{A} \leq \mathcal{B}$	$O(n^2)$ [9]	$O(1)$ [11, 24]	$O(n^2)$ [9, 23]	$O(1)$ [15]	$O(1)$ [21]	$O(1)$ [15]
$\mathcal{A} \geq \mathcal{B}$		NP-hard [8]	$O(1)$ [6]	NP-hard [8, 22]	$O(n)$ [15, 19, 20]	$O(n)$ [21]
$\mathcal{A} \leq \mathcal{C}$			$O(n^2)$ [23]	$O(1)$ [11, 24]	$O(1)$ [6]	$O(1)$ [6]
$\mathcal{A} \geq \mathcal{C}$				NP-hard [22]	$O(1)$ [6]	$O(1)$ [6]
$\mathcal{B} \leq \mathcal{C}$					NP-hard (new)	$O(n^2)$ [16, 6]
$\mathcal{B} \geq \mathcal{C}$						$O(n^4)$ [6]

Given a relaxed Supnick matrix with indeterminate diagonal elements, it is not always possible to define the diagonal elements so that the resulting matrix is a Monge matrix, as the following example shows:

$$\begin{pmatrix} \times & 1 & 0 & 1 & 5 \\ 1 & \times & 0 & 0 & 3 \\ 0 & 0 & \times & 1 & 1 \\ 1 & 0 & 1 & \times & 0 \\ 5 & 3 & 1 & 0 & \times \end{pmatrix}$$

Computational complexity of the TSP on the described matrix classes, and other classes of similar type, can be studied systematically by considering *four-point conditions*. Let i, j, k, l be four points with $1 \leq i < j < k < l \leq n$. A symmetric distance matrix for these points contains six entries above the main diagonal, which correspond to the six edges connecting these points. It is possible to form three pairs of non-incident edges: $\{(i, j), (k, l)\}$, $\{(i, k), (j, l)\}$, $\{(i, l), (j, k)\}$. We denote the combined lengths of these pairs by $\mathcal{A}, \mathcal{B}, \mathcal{C}$, respectively:

$$\mathcal{A} = c_{ij} + c_{kl} \quad \mathcal{B} = c_{ik} + c_{jl} \quad \mathcal{C} = c_{il} + c_{jk}$$

A *four-point condition* is an inequality relation among the values $\mathcal{A}, \mathcal{B}, \mathcal{C}$, which has to be satisfied for all possible choices of indices i, j, k, l with $1 \leq i < j < k < l \leq n$. Using this notation, Supnick matrices are defined as matrices satisfying conditions $\mathcal{A} \leq \mathcal{B}, \mathcal{B} \leq \mathcal{C}$, Demidenko matrices as matrices satisfying condition $\mathcal{A} \leq \mathcal{B}$, Van der Veen matrices as matrices satisfying conditions $\mathcal{A} \leq \mathcal{C}$, and relaxed Supnick matrices as matrices satisfying condition $\mathcal{B} \leq \mathcal{C}$. Nearly all possible four-point conditions and their pairwise combinations have been

classified in [6] (see Table 1) according to the polynomial solvability or NP-hardness of the arising TSP. The only gap that has remained in this classification is the TSP on relaxed Supnick matrices. In the following section we show that this problem is NP-hard, thus making a final point in the classification of four-point conditions for symmetric TSP.

2 The TSP on Relaxed Supnick Matrices

As before, we assume that all the considered matrices are symmetric.

Theorem 1. *The TSP on a relaxed Supnick matrix is NP-hard.*

Proof. The proof is by reduction from the Hamiltonian cycle problem in grid graphs [13]. We follow the definitions from [13]. Let G^∞ be the infinite graph, whose vertex set consists of all the integer points in the plane, and in which two vertices are connected, if and only if the Euclidean distance between them is equal to one. A grid graph is a finite node-induced subgraph of G^∞ . It is shown in [13] that the Hamiltonian cycle problem in a grid graph is NP-hard.

Given a grid graph on n nodes, we will construct an $n \times n$ relaxed Supnick matrix with non-negative entries, such that there exists a Hamiltonian cycle in the graph, if and only if the optimal TSP tour on the corresponding relaxed Supnick matrix has zero length.

Consider an arbitrary grid graph G on n nodes. We may assume that G is connected (otherwise, it is guaranteed not to contain a Hamiltonian cycle). First, we embed G in a square grid of size $m \times m$, where m is sufficiently large. Due to the connectivity of G , we may assume $m \leq n$. We then extend the square grid to a parallelogram grid, and number the nodes as shown in Figure 1. The upper-left point in the square grid will be numbered m , the two neighbouring points $2m - 1$ and $2m$, etc. Let $N = m^2 + m(m - 1)$ be the number of points in the parallelogram grid. We then construct an $N \times N$ relaxed Supnick matrix $C^{(N)}$ as follows. For two neighbouring nodes $i, j, i < j$, in the parallelogram grid, we define $c_{ij}^{(N)} = 0$, if both i and j belong to G . Notice that the numbering of points in the parallelogram grid is chosen so that all these entries are placed on two

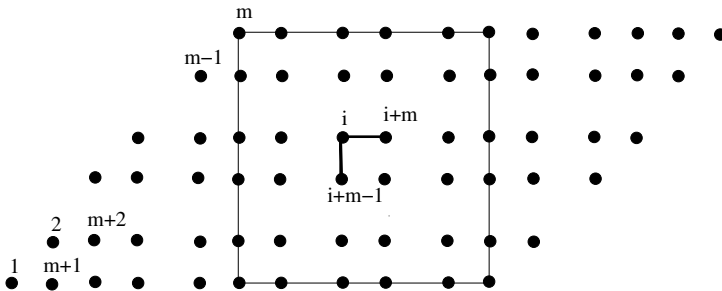


Fig. 1. Embedding a square grid in a parallelogram

adjacent diagonals: $j - i \in \{m - 1, m\}$. For all the remaining entries in these two diagonals, we define $c_{ij}^{(N)} = 1$. Further, we define $c_{1j}^{(N)} = 1$ for $j = 1, \dots, m - 1$, and $c_{iN}^{(N)} = 1$ for $i = N - m + 2, \dots, N$.

Notice that so far there exist no four defined entries in the upper triangular part of $C^{(N)}$ that would form an inequality from Definition 2, therefore none of the one-sided Monge conditions are violated. We keep filling in the upper triangular part of the matrix, respecting the Monge conditions. We define

$$\begin{aligned}
 c_{ij}^{(N)} &= \max\{c_{i-1,j}^{(N)} + c_{i,j+1}^{(N)} - c_{i-1,j+1}^{(N)}, 1\} & j - i = 1, \dots, m - 2 \\
 c_{ij}^{(N)} &= \max\{c_{i,j-1}^{(N)} + c_{i+1,j}^{(N)} - c_{i+1,j-1}^{(N)}, 1\} & j - i = m + 1, \dots, N - 1
 \end{aligned}$$

for each element, as soon as the right-hand side of this element’s definition has itself been defined. We always have $c_{ij}^{(N)} \geq 1$ unless i, j are neighbouring nodes in G , and the one-sided Monge condition is always preserved. The lower triangular part of $C^{(N)}$ is defined by symmetry. The construction of the relaxed Supnick matrix $C^{(N)}$ is now completed.

It can be easily checked that, given a relaxed Supnick matrix C , the symmetric matrix obtained by deleting a row and a corresponding column from C is still relaxed Supnick. By induction, the same is true after deleting an arbitrary subset of rows and corresponding columns. By deleting all rows and columns from $C^{(N)}$, except those indexed by points in the original grid graph G , we obtain an $n \times n$ relaxed Supnick matrix $C^{(n)}$. Clearly, there exists a Hamiltonian cycle in graph G , if and only if the optimal TSP tour on the matrix $C^{(n)}$ has zero length.

In order to prove that the above problem reduction is polynomial, it only remains to show that the growth of individual matrix elements is appropriately bounded. Observe that each new element created in $C^{(N)}$ cannot exceed the largest existing element by more than a factor of 2. Consequently, the largest element in $C^{(N)}$ has value¹ at most $2^{N^2} \leq 2^{(2m^2)^2} = 2^{4m^4} \leq 2^{4n^4}$, and can therefore be represented by a number of bits that is polynomial in n . Hence, even if our computational model does not support unit-cost arithmetic on arbitrarily long integers, arithmetic operations on matrix elements can be emulated bitwise in polynomial time. The reduction is completed. □

3 Relaxed Supnick Matrices and Exponential Neighbourhoods

The research into polynomially solvable cases for intractable problems is partly motivated by the hope to identify new approaches to constructing general-purpose heuristics. One of the by-products of this research are exponential neighbourhoods (see surveys [1, 7]), which are being intensively studied in relation to

¹ Much tighter estimates are possible. However, this crude bound is sufficient for our proof.

local search algorithms. Among new families of exponential neighbourhoods presented in [6], there is a neighbourhood of *strongly balanced permutations*, which are related to relaxed Supnick matrices. To justify the introduction of this neighbourhood, the authors of [6] considered a special subclass of relaxed Supnick matrices.

Definition 4. A relaxed Supnick matrix $C = (c_{ij})$ is strong, if

$$c_{tp} - c_{tz} - c_{kp} \leq c_{sy} - c_{ky} - c_{sz} \quad t < k < s < p < z < y$$

It can be shown that the inequalities in Definition 4 are equivalent to the system of inequalities

$$c_{ij} - c_{i,j+1} - c_{i+1,j} \leq c_{j-1,N} - c_{i+1,N} - c_{j-1,j} \quad i < j$$

and can therefore be checked in time $O(n^2)$.

As shown in [6], an optimal tour for the TSP on a strong relaxed Supnick matrix can be found in the set of so-called strongly balanced tours. We first give some preliminary definitions. An index $i \in \{1, \dots, n\}$ is a *peak* of a permutation τ , if $i > \max\{\tau^{-1}(i), \tau(i)\}$, and a *valley*, if $i < \min\{\tau^{-1}(i), \tau(i)\}$. An index which is neither peak nor valley is called *intermediate*. Informally, in a strongly balanced tour, all intermediate nodes are “evenly spread” on the slopes between peaks and valleys. We now give a formal definition.

We will consider partially constructed tours on the sets of indices $\{1, 2, 3, \dots, m - 1, m\}$ with $m = 1, 2, \dots, n$. For a fixed m , a partially constructed tour will consist of a set of finite index sequences; indices from each sequence are placed in the tour in consecutive positions. We refer to each of these sequences $\langle i_1, \dots, j_1 \rangle$ as *fragment* $[i_1, j_1]$, stressing that i_1 is the initial, and j_1 is the final element in the corresponding sequence. Notice that it is not necessary for a fragment $[i_1, j_1]$ with $i_1 < j_1$ to contain, for example, $i_1 + 1$. For a one-element fragment we use the notation $[i, i]$. For example, if we start with a tour where 1 and 2 are two valleys, we can represent this initial tour by two fragments $[1, 1]$ and $[2, 2]$. Mutual placement of fragments is not fixed, i.e. they can be permuted. The fragments can also be reversed, i.e. fragment $[i, j]$ can be replaced by fragment $[j, i]$.

Definition 5 ([6]). A tour is strongly balanced, if it can be constructed as follows. Start with an initial tour $[1, 1]$ and repeat the following step for $m = 2, \dots, n - 1$:

Given a partial tour on the set of indices $\{1, \dots, m - 1\}$, the tour is represented by fragments $[i_1, j_2], [i_2, j_2], \dots, [i_p, j_p]$. Let

$$\begin{aligned} i_{min1} &= \min\{i_1, j_1, i_2, j_2, \dots, i_p, j_p\} \\ i_{min2} &= \min\{i_1, j_1, i_2, j_2, \dots, i_p, j_p\} \setminus \{i_{min1}\} \\ i_{min3} &= \min\{i_1, j_1, i_2, j_2, \dots, i_p, j_p\} \setminus \{i_{min1}, i_{min2}\} \end{aligned}$$

Add index m to the partial tour by choosing one of the options below:

- m is placed as a new valley; this creates a new fragment $[m, m]$;
- m is placed as an intermediate index adjacent to i_{min1} ; fragment $[i_{min1}, s]$ in the partially constructed tour is replaced by the new fragment $[m, s]$;
- m is placed as a new peak merging two fragments; m is adjacent to i_{min1} and to either i_{min2} or to i_{min3} . In the first case, the fragments $[i_{min1}, j]$ and $[i_{min2}, s]$ are merged into $[j, s]$; in the second case, the fragments $[i_{min1}, i_{min2}]$ and $[i_{min3}, s]$ are merged into $[i_{min2}, s]$.

The final node n can only be added to a partial tour consisting of one fragment.

As an example, the reader can check that the tour

$$\langle 1, 4, 10, 6, 2, 7, 12, 9, 3, 8, 11, 5, 1 \rangle$$

with the intermediate nodes 4, 5, 6, 7, 8, 9 evenly spread on the slopes, is a strongly balanced tour.

Proposition 1 ([6]). *An optimal tour for the TSP with a strong reduced Supnick matrix can be found in the set of strongly balanced tours.*

Despite the relatively simple structure of strongly balanced tours, the problem of finding an optimal tour in this set appears to be difficult. To avoid the difficulties, the authors of [6] have further restricted the special class of matrices, and identified a subset of strongly balanced tours, where an optimal tour can be found in polynomial time.

Proposition 2 ([6]). *Consider strongly balanced tours for which the maximum number of fragments in the construction of Definition 5 is bounded by a constant. An optimal tour in this special subset of strongly balanced tours can be found in polynomial time.*

We can now explain why the problem of finding an optimal strongly balanced tour is difficult for an unbounded number of fragments.

Theorem 2. *The TSP on a strong relaxed Supnick matrix is NP-hard.*

Proof. The proof is similar to the proof of Theorem 1. We define the initial values $c_{ij}^{(N)}$ in two adjacent diagonals: $j - i \in \{m - 1, m\}$ exactly as before. The only difference is in the way we define the remaining entries. To ensure that the matrix is a strong relaxed Supnick matrix, it is sufficient to define

$$c_{ij}^{(N)} = \max\{c_{j,i-1}^{(N)} + c_{j+1,i-1}^{(N)} - c_{j+1,i-1}^{(N)}, \\ c_{j,i-1}^{(N)} + c_{i-1,i}^{(N)} + c_{j+1,N}^{(N)} - c_{i-1,N}^{(N)} - c_{j+1,i-1}^{(N)}, 1\} \quad j < j + m < i$$

and

$$c_{ip}^{(N)} = \max\{c_{i-1,p}^{(N)} + c_{i,i+1}^{(N)} - c_{i-1,p+1}^{(N)}, \\ c_{i-1,p}^{(N)} + c_{p,p+1}^{(N)} + c_{iN}^{(N)} - c_{pN}^{(N)} - c_{i-1,p+1}^{(N)}, 1\} \quad i < p < i + m$$

The rest of the proof is identical to the proof of Theorem 1, with the only difference that the largest element in $C^{(N)}$ has now value at most $3^{N^2} \leq 3^{4n^4}$. \square

Corollary 1. *The problem of finding an optimal strongly balanced tour is NP-hard.*

The two statements above justify the introduction of a special subset of strongly balanced tours (Proposition 2), where the optimal permutation can be found in polynomial time.

4 Conclusion

We have shown that the TSP on a relaxed Supnick matrix, which can also be viewed as a one-sided Monge matrix, is NP-hard. This completes the classification of all possible four-point conditions for symmetric TSP [6]. Our results justify imposing further restrictions on relaxed Supnick matrices in order to identify new polynomially solvable cases of the TSP.

We hope that the constructions in this note, which reduce the TSP on a grid graph to the TSP with a special matrix, in combination with the special polynomially solvable case described in Proposition 2 (see [6] for further details), may lead to identifying special types of grid graphs, where the optimal tour can be found in polynomial time. Some polynomially solvable cases of the TSP on grid graphs have already been identified in [2, 18].

References

1. Ahuja, R.K., Ergun, Ö., Orlin, J.B., Punnen, A.: A survey of very large-scale neighborhood search techniques. *Discrete Applied Mathematics*, **123** (2002) 75–102
2. Arkin, E.M., Bemder, M.A., Demaine, E., Fekete, S.P., Mitchell, J.S., Sthia, S.: Optimal covering tours with turn costs. In *Proc. 13th ACM-SIAM Symposium on Discrete Algorithms, SODA01* (2001) 138–147
3. Burkard, R.E., Deineko, V.G., van Dal, R., van der Veen, J.A.A., Woeginger, G.J.: Well-solvable special cases of the TSP: A survey. *SIAM Review*, **40**, 3 (1998) 496–546
4. Burkard, R.E., Deineko, V.G.: On the traveling salesman problem with a relaxed Monge matrix. *Information Processing Letters*, **67** (1998) 231–237
5. Burkard, R.E., Klinz, B., Rudolf, R.: Perspectives of Monge Properties in Optimization. *Discrete Applied Mathematics*, **70** (1996) 95–161
6. Deineko, V.G., Klinz, B., Woeginger, G.J.: Four point conditions for symmetric TSP and exponential neighbourhoods. In *Proc. 17th ACM-SIAM Symposium on Discrete Algorithms, SODA 06* (2006) 544–553
7. Deineko, V.G., Woeginger, G.J.: A study of exponential neighborhoods for the travelling salesman problem and for the quadratic assignment problem. *Mathematical Programming* **87** (2000) 519–542
8. Deineko, V.G., Woeginger, G.J.: The maximum travelling salesman problem on symmetric Demidenko matrices. *Discrete Applied Mathematics*, **99** (2000) 413–425

9. Demidenko, V.M.: A special case of traveling salesman problems. *Izv. Akad. Nauk. BSSR, Ser. Fiz.-mat. Nauk*, 5 (1976) 28–32 (in Russian)
10. Garey, M.R., Johnson, D.S.: *Computers and intractability*. Freeman and Co., San Francisco (1979)
11. Gilmore, P.C., Lawler, E.L., Shmoys, D.B.: Well-solved special cases. Chapter 7 in [17] (1985) 207–249
12. Gutin, G., Punnen, A.P. (eds.): *The travelling salesman problem and its variations*. Kluwer Academic Publishers (2002)
13. Itai, A., Papadimitiou, C.H., Szwarcfiter, J.L.: Hamiltonian paths in grid graphs. *SIAM Journal on Computing*, 11, 4 (1982) 676–686
14. Kabadi, S.N.: Polynomially solvable cases of the TSP. Chapter 11 in [12] (2002) 489–583
15. Kalmanson, K.: Edgeconvex circuits and the traveling salesman problem. *Canadian Journal of Mathematics*, 27 (1975) 1000–1010
16. Lawler, E.L.: A solvable case of the traveling salesman problem. *Mathematical Programming*, 1 (1971) 267–269
17. Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G., Shmoys, D.B.: *The Traveling Salesman Problem*. Wiley, Chichester (1985)
18. Lenhart, W., Umans, U.: Hamiltonian cycles in solid grid graphs. In *Proceedings of 38 Annual Symposium on Foundations of Computer Science, FOCS '97* (1997) 496–596
19. Quintas, L.V., Supnick, F.: Extreme Hamiltonian Circuits: Resolution of the convex-odd case. *Proc. Amer. Math. Soc.*, 15 (1964) 454–459
20. Quintas, L.V., Supnick, F.: Extreme Hamiltonian Circuits: Resolution of the convex-even case. *Proc. Amer. Math. Soc.*, 16 (1965) 1058–1061
21. Supnick, F.: Extreme Hamiltonian lines. *Annals of Math.*, 66 (1957) 179–201
22. Steiner, G., Xue, X.: The maximum traveling salesman problem on van der Veen matrices. *Discrete Applied Mathematics*, 146 (2005) 1–2
23. van der Veen, J.A.A.: A new class of pyramidally solvable symmetric traveling salesman problems. *SIAM J. Discr. Math.*, 7 (1994) 585–592
24. Volodarskiy, Y.M., Gabovich, E.Y., Zacharin, A.Y.: A solvable case in discrete programming. *Izv. Akad. Nauk. SSSR, Ser. Techn. Kibernetika*, 1 (1976) 34–44 (in Russian). English translation in *Engineering Cybernetics*, 14 (1977) 23–32

On Direct Methods for Lexicographic Min-Max Optimization*

Włodzimierz Ogryczak and Tomasz Śliwiński

Warsaw University of Technology, Institute of Control & Computation Engineering,
00-665 Warsaw, Poland
{wogrycza, tsliwins}@ia.pw.edu.pl

Abstract. The approach called the Lexicographic Min-Max (LMM) optimization depends on searching for solutions minimal according to the lex-max order on a multidimensional outcome space. LMM is a refinement of the standard Min-Max optimization, but in the former, in addition to the largest outcome, we minimize also the second largest outcome (provided that the largest one remains as small as possible), minimize the third largest (provided that the two largest remain as small as possible), and so on. The necessity of point-wise ordering of outcomes within the lexicographic optimization scheme causes that the LMM problem is hard to implement. For convex problems it is possible to use iterative algorithms solving a sequence of properly defined Min-Max problems by eliminating some blocked outcomes. In general, it may not exist any blocked outcome thus disabling possibility of iterative Min-Max processing. In this paper we analyze two alternative optimization models allowing to form lexicographic sequential procedures for various nonconvex (possibly discrete) LMM problems. Both the approaches are based on sequential optimization of directly defined artificial criteria. The criteria can be introduced into the original model with some auxiliary variables and linear inequalities thus the methods are easily implementable.

1 Lexicographic Min-Max

There are several multiple criteria decision problems where the Pareto-optimal solution concept is not powerful enough to resolve the problem since the equity or fairness among uniform individual outcomes is an important issue [10, 11, 17]. Uniform and equitable outcomes arise in many dynamic programs where individual objective functions represent the same outcome for various periods [9]. In the stochastic problems uniform objectives may represent various possible values of the same nondeterministic outcome ([15] and references therein). Moreover, many modeling techniques for decision problems first introduce some uniform objectives and next consider their impartial aggregations. The most direct models with uniform equitable criteria are related to the optimization of systems

* The research was supported by the Ministry of Science and Information Society Technologies under grant 3T11C 005 27 “Models and Algorithms for Efficient and Fair Resource Allocation in Complex Systems.”

which serve many users. For instance, efficient and fair way of distribution of network resources among competing demands becomes a key issue in computer networks [5] and the telecommunication networks design, in general [20].

The generic decision problem we consider may be stated as follows. There is given a set I of m clients (users, services). There is also given a set Q of feasible decisions. For each service $j \in I$ a function $f_j(\mathbf{x})$ of the decision \mathbf{x} is defined. This function, called the individual objective function, measures the outcome (effect) $y_j = f_j(\mathbf{x})$ of the decision for client j . An outcome can be measured (modeled) as service time, service costs, service delays as well as in a more subjective way as individual utility (or disutility) level. In typical formulations a smaller value of the outcome means a better effect (higher service quality or client satisfaction). Therefore, without loss of generality, we can assume that each individual outcome y_i is to be minimized.

The Min-Max solution concept depends on optimization of the worst outcome

$$\min_{\mathbf{x}} \{ \max_{j=1, \dots, m} f_j(\mathbf{x}) : \mathbf{x} \in Q \}$$

and it is regarded as maintaining equity. Indeed, for a simplified resource allocation problem $\min\{\max_j y_j : \sum_j y_j \leq b\}$ the Min-Max solution takes the form $\bar{y}_j = b/m$ for all $j \in I$ thus meeting the perfect equity. In the general case with possibly more complex feasible set structure this property is not fulfilled. Actually, the distribution of outcomes may make the Min-Max criterion partially passive when one specific outcome is relatively large for all the solutions. For instance, while allocating clients to service facilities, such a situation may be caused by existence of an isolated client located at a considerable distance from all facilities. Minimization of the maximum distance is then reduced to that single isolated client leaving other allocation decisions unoptimized. This is a clear case of inefficient solution where one may still improve other outcomes while maintaining fairness (equitability) by leaving at its best possible value the worst outcome.

The Min-Max solution may be lexicographically regularized according to the Rawlsian principle of justice [22]. Applying the Rawlsian approach, any two states should be ranked according to the accessibility levels of the least well-off individuals in those states; if the comparison yields a tie, the accessibility levels of the next-least well-off individuals should be considered, and so on. Formalization of this concept leads us to the lexicographic Min-Max optimization. Let $\langle \mathbf{a} \rangle = (a_{(1)}, a_{(2)}, \dots, a_{(m)})$ denote the vector obtained from \mathbf{a} by rearranging its components in the non-increasing order. That means $a_{(1)} \geq a_{(2)} \geq \dots \geq a_{(m)}$ and there exists a permutation π of set I such that $a_{(i)} = a_{\pi(i)}$ for $i \in I$. Comparing lexicographically such ordered vectors $\langle \mathbf{y} \rangle$ one gets the so-called lex-max order. The general problem we consider depends on searching for the solutions that are minimal according to the lex-max order:

$$\text{lex min}_{\mathbf{x}} \{ (\theta_1(\mathbf{f}(\mathbf{x})), \dots, \theta_m(\mathbf{f}(\mathbf{x}))) : \mathbf{x} \in Q \} \quad \text{where } \theta_j(\mathbf{y}) = y_{(j)} \quad (1)$$

The lexicographic Min-Max under consideration is related to the problems with outcomes being minimized. Similar consideration of the maximization problems

leads to the lexicographic Max-Min solution concept. Obviously, all the results presented further for the lexicographic Min-Max can be adjusted to the lexicographic Max-Min while preserving assumption that the outcomes are ordered from the worst one to the best one.

The lexicographic Min-Max solution is known in game theory as the nucleolus of a matrix game. It originates from an idea [6] to select from the optimal strategy set those which allow one to exploit mistakes of the opponent optimally. It has been later refined to the formal nucleolus definition [21]. The concept was early considered in the Tschebyscheff approximation [23] as a refinement taking into account the second largest deviation, the third one and further to be hierarchically minimized. Similar refinement of the fuzzy set operations has been recently analyzed [7]. Within the telecommunications or network applications the lexicographic Max-Min approach has appeared already in [2] and now under the name Max-Min Fairness (MMF) is treated as one of the standard fairness concepts [16, 20]. The LMM approach has been used for general linear programming multiple criteria problems [1, 12], as well as for specialized problems related to (multiperiod) resource allocation [9, 11].

Note that the lexicographic minimization in the LMM is not applied to any specific order of the original criteria. Nevertheless, in the case of linear programming (LP) problems (or generally convex optimization), there exists a dominating objective function which is constant (blocked) on the entire optimal set of the Min-Max problem [12]. Hence, having solved the Min-Max problem, one may try to identify the blocked objective and eliminate it to formulate a new restricted Min-Max problem on the former optimal set. Therefore, the LMM solution to LP problems can be found by the sequential Min-Max optimization with elimination of the blocked outcomes.

The LMM approach has been considered also for various discrete optimization problems [3, 4, 8] including the location-allocation ones [14]. In discrete models, due to the lack of convexity there may not exist any blocked outcome [13] thus disabling possibility of the sequential Min-Max algorithm. In this paper we analyze capabilities of an effective use of earlier developed ordered cumulated outcomes methodology [17, 18, 19] to solve the LMM problem by sequential optimization of directly defined criteria. We develop and analyze two alternative approaches allowing to form lexicographic sequential procedures for various non-convex (possibly discrete) LMM problems. Both the approaches are based on criteria directly introduced with some LP expansion of the original model.

2 Direct Models

2.1 Ordered Outcomes

The ordered outcomes $y_{\langle k \rangle}$ used in definition of the LMM solution concept can be expressed with a direct formula, although requiring the use of integer variables [24]. Namely, for any $k = 1, 2, \dots, m$ the following formula is valid:

$$y_{\langle k \rangle} = \min_{t_k, z_{kj}} \{t_k : t_k - y_j \geq -Mz_{kj}, z_{kj} \in \{0, 1\} \forall j, \sum_{j=1}^m z_{kj} \leq k - 1\}$$

where M is a sufficiently large constant (larger than the span of individual outcomes y_j) which allows one to enforce inequality $t_k \geq y_j$ for $z_{kj} = 0$ while ignoring it for $z_{kj} = 1$. Note that for $k = 1$ all binary variables z_{1j} are forced to 0 thus reducing the optimization in this case to the standard LP model. However, for any other $k > 1$ all m binary variables z_{kj} are an important part of the model. Nevertheless, with the use of auxiliary integer variables, any LMM problem (either convex or non-convex) can be formulated as the standard lexicographic minimization with directly given objective functions

$$\begin{aligned} \text{lex min}_{\mathbf{x}, t_k, z_{kj}} (t_1, t_2, \dots, t_m) \text{ s.t. } \mathbf{x} \in Q, \quad & \sum_{j=1}^m z_{kj} \leq k - 1 \quad \forall k \\ & t_k - f_j(\mathbf{x}) \geq -Mz_{kj}, \quad z_{kj} \in \{0, 1\} \quad \forall j, k \end{aligned} \tag{2}$$

Let us consider cumulated criteria $\bar{\theta}_k(\mathbf{y}) = \sum_{i=1}^k y_{(i)}$ expressing, respectively: the worst (largest) outcome, the total of the two worst outcomes, the total of the three worst outcomes, etc. When normalized by k the quantities $\mu_k(\mathbf{y}) = \bar{\theta}_k(\mathbf{y})/k$ can be interpreted as the worst conditional means [17]. Within the lexicographic optimization a cumulation of criteria does not affect the optimal solution. Hence, the LMM problem can be formulated as the standard lexicographic minimization with cumulated ordered outcomes as objective functions

$$\text{lex min}_{\mathbf{x}} \{(\bar{\theta}_1(\mathbf{f}(\mathbf{x})), \dots, \bar{\theta}_m(\mathbf{f}(\mathbf{x}))) : \mathbf{x} \in Q\}$$

where $\bar{\theta}_k(\mathbf{y}) = \sum_{i=1}^k y_{(i)}$. This simplifies dramatically the optimization problem since quantities $\bar{\theta}_k(\mathbf{y})$ can be optimized without use of any integer variables. First, let us notice that for any given vector \mathbf{y} , the cumulated ordered value $\bar{\theta}_k(\mathbf{y})$ can be found as the optimal value of the following LP problem:

$$\bar{\theta}_k(\mathbf{y}) = \max_{u_{kj}} \left\{ \sum_{j=1}^m y_j u_{kj} : \sum_{j=1}^m u_{kj} = k, \quad 0 \leq u_{kj} \leq 1 \quad \forall j \right\} \tag{3}$$

The above problem is an LP for a given outcome vector \mathbf{y} while it becomes non-linear for \mathbf{y} being a vector of variables. This difficulty can be overcome by taking advantage of the LP dual to (3). Introducing dual variable t_k corresponding to the equation $\sum_{j=1}^m u_{kj} = k$ and variables d_{kj} corresponding to upper bounds on u_{kj} one gets the following LP dual of problem (3):

$$\bar{\theta}_k(\mathbf{y}) = \min_{t_k, d_{kj}} \left\{ kt_k + \sum_{j=1}^m d_{kj} : t_k + d_{kj} \geq y_j, \quad d_{kj} \geq 0 \quad \forall j \right\} \tag{4}$$

Due to the duality theory, for any given vector \mathbf{y} the cumulated ordered coefficient $\bar{\theta}_k(\mathbf{y})$ can be found as the optimal value of the above LP problem.

It follows from (4) that $\bar{\theta}_k(\mathbf{f}(\mathbf{x})) = \min \{kt_k + \sum_{j=1}^m (f_j(\mathbf{x}) - t_k)_+ : \mathbf{x} \in Q\}$, where $(\cdot)_+$ denotes the nonnegative part of a number and t_k is an auxiliary (unbounded) variable. This is equivalent to the computational formulation of the

k -centrum model introduced in [19]. Hence, formula (4) provides an alternative proof of that formulation.

Following formula (4), the following assertion is valid for any LMM problem.

Theorem 1. *Every optimal solution to the LMM problem (1) can be found as an optimal solution to a standard lexicographic optimization problem with predefined linear criteria:*

$$\begin{aligned} \text{lex } \min_{\mathbf{x}, t_k, d_{kj}} & \left[t_1 + \sum_{j=1}^m d_{1j}, 2t_2 + \sum_{j=1}^m d_{2j}, \dots, mt_m + \sum_{j=1}^m d_{mj} \right] \\ \text{s.t. } \mathbf{x} \in Q, & \quad t_k + d_{kj} \geq f_j(\mathbf{x}), \quad d_{kj} \geq 0 \quad \forall j, k \end{aligned} \tag{5}$$

This direct lexicographic formulation remains valid for nonconvex (e.g. discrete) feasible sets Q , where the standard sequential approaches [11, 12] are not applicable [14]. Note that model (5) does not use integer variables and it can be considered as an LP expansion of the original Min-Max problem. Thus, this model preserves the problem’s convexity if the original problem is defined with a convex feasible set Q and a linear objective functions f_j . The size of the problem is quadratic with respect to the number of outcomes ($m^2 + m$ auxiliary variables and m^2 constraints).

2.2 Ordered Values

For some specific classes of discrete, or rather combinatorial, optimization problems, one may take advantage of the finiteness of the set of all possible values of functions f_j on the finite set of feasible solutions. The ordered outcome vectors may be treated as describing a distribution of outcomes generated by a given decision \mathbf{x} . In the case when there exists a finite set of all possible outcomes of the individual objective functions (or the set of outcome values can be restricted to its finite approximation, i.e. with fuzzy modeling), one can directly describe the distribution of outcomes with frequencies of outcomes. Let $V = \{v_1, v_2, \dots, v_r\}$ (where $v_1 > v_2 > \dots > v_r$) denote the set of all attainable outcomes (all possible values of the individual objective functions f_j for $\mathbf{x} \in Q$). We introduce integer functions $h_k(\mathbf{y})$ ($k = 1, 2, \dots, r$) expressing the number of values v_k in the outcome vector \mathbf{y} . Having defined functions h_k we can introduce cumulative distribution functions $\bar{h}_k(\mathbf{y}) = \sum_{l=1}^k h_l(\mathbf{y})$ where $\bar{h}_r(\mathbf{y}) = m$ for any outcome vector. Function \bar{h}_k expresses the number of outcomes larger or equal to v_k . Since we want to minimize all the outcomes, we are interested in the minimization of all functions \bar{h}_k for $k = 1, 2, \dots, r - 1$. Indeed, the LMM solution concept can be expressed in terms of the standard lexicographic minimization problem with objectives $\bar{h}_k(\mathbf{f}(\mathbf{x}))$ [13].

Theorem 2. *In the case of finite outcome set $\mathbf{f}(Q) = V^m$, the LMM problem (1) is equivalent to the standard lexicographic optimization problem with $r - 1$ criteria:*

$$\text{lex } \min_{\mathbf{x}} \{ (\bar{h}_1(\mathbf{f}(\mathbf{x})), \dots, \bar{h}_{r-1}(\mathbf{f}(\mathbf{x}))) : \mathbf{x} \in Q \} \tag{6}$$

Unfortunately, for functions \bar{h}_k there is no simple analytical formula allowing to minimize them without use of some auxiliary integer variables. This difficulty can be overcome by taking advantages of possible weighting and cumulating criteria in lexicographic optimization. Namely, for any positive weights w_i the lexicographic optimization

$$\text{lex min}_{\mathbf{x}} \{ (w_1 \bar{h}_1(\mathbf{f}(\mathbf{x})), w_1 \bar{h}_1(\mathbf{f}(\mathbf{x})) + w_2 \bar{h}_2(\mathbf{f}(\mathbf{x})), \dots, \sum_{i=1}^{r-1} w_i \bar{h}_i(\mathbf{f}(\mathbf{x}))) : \mathbf{x} \in Q \}$$

is equivalent to (6). Let us cumulate vector $\bar{\mathbf{h}}(\mathbf{y})$ weights $w_i = v_i - v_{i+1}$ to get

$$\hat{h}_k(\mathbf{y}) = \sum_{i=1}^{k-1} (v_i - v_{i+1}) \bar{h}_i(\mathbf{y}) = \sum_{i=1}^{k-1} (v_i - v_k) h_i(\mathbf{y}) = \sum_{j=1}^m (y_j - v_k)_+$$

where quantities $\hat{h}_k(\mathbf{f}(\mathbf{y}))$ for $k = 2, 3, \dots, r$ represent the total exceed of outcomes over the corresponding values v_k . Due to the use of positive weights $w_i > 0$, the lexicographic problem (6) is equivalent to the lexicographic minimization

$$\text{lex min}_{\mathbf{x}} \{ (\hat{h}_2(\mathbf{f}(\mathbf{x})), \dots, \hat{h}_r(\mathbf{f}(\mathbf{x}))) : \mathbf{x} \in Q \}$$

Moreover, criteria defined this way are piecewise linear convex functions [13] which allow to compute them directly by the minimization:

$$\hat{h}_k(\mathbf{y}) = \min_{v_k, h_{kj}} \left\{ \sum_{j=1}^m h_{kj} : h_{kj} \geq y_j - v_k, h_{kj} \geq 0 \forall j \right\}$$

Therefore, the following assertion is valid for any LMM problem.

Theorem 3. *In the case of finite outcome set $\mathbf{f}(Q) = V^m$, every optimal solution to the LMM problem (1) can be found as an optimal solution to a standard lexicographic optimization problem with predefined linear criteria:*

$$\begin{aligned} \text{lex min}_{\mathbf{x}, v_k, h_{kj}} & \left[\sum_{j=1}^m h_{2j}, \sum_{j=1}^m h_{3j}, \dots, \sum_{j=1}^m h_{rj} \right] \\ \text{s.t. } \mathbf{x} \in Q, & \quad h_{kj} \geq f_j(\mathbf{x}) - v_k, h_{kj} \geq 0 \quad \forall j, k \end{aligned} \tag{7}$$

Formulation (7) does not use integer variables and can be considered as an LP expansion of the original problem. Thus, this model preserves the problem's convexity if the original problem is defined with a convex feasible set Q and objective functions f_j . The size of the problem depends on the number of different outcome values. For many models with not too large number of outcome values, the problem can easily be solved directly and even for convex problems such an approach may be more efficient than the sequential algorithms. Note that in many problems of systems optimization the objective functions express the quality of service and one can easily consider a limited finite scale (grid) of the corresponding outcome values (possibly fuzzy values).

3 Computational Experiments

We have run initial tests to analyze the computational performances of both direct models for the LMM problem. For this purpose we have solved randomly generated (discrete) location problems defined as follows. There is given a set of m clients. There is also given a set of n potential locations for the facilities, in particular, we considered all points representing the clients as valid potential locations ($n = m$). Further, the number (or the maximal number) p of facilities to be located is given ($p \leq n$). The main decisions to be made in the location problem can be described with the binary variables: x_i equal to 1 if location i is to be used and equal to 0 otherwise ($i = 1, 2, \dots, n$). The allocation decisions are modeled with the additional allocation variables: x'_{ij} equal to 1 if location i is used to service client j and equal to 0 otherwise ($i = 1, 2, \dots, n; j = 1, 2, \dots, m$).

$$\begin{aligned} \sum_{i=1}^n x_i = p, \quad \sum_{i=1}^n x'_{ij} = 1 \quad & j = 1, 2, \dots, m \\ x'_{ij} \leq x_j, \quad x_j, x'_{ij} \in \{0, 1\} \quad & i = 1, 2, \dots, n; j = 1, 2, \dots, m \end{aligned} \tag{8}$$

For each client j a function $f_j(\mathbf{x})$ of the location pattern \mathbf{x} has been defined to measure the outcome (effect) of the location pattern for client j . Individual objective functions f_j depend on effects of several allocations decisions. It means they depend on allocation effect coefficients $d_{ij} > 0$ ($i = 1, \dots, m; j = 1, \dots, n$), called hereafter simply distances as they usually express the distance (or travel time) between location i and client j . For the standard uncapacitated location problem it is assumed that all the potential facilities provide the same type of service and each client is serviced by the nearest located facility. With the explicit use of the allocation variables and the corresponding constraints the individual objective functions f_j can be written in the linear form: $f_j(\mathbf{x}) = \sum_{i=1}^n d_{ij} x'_{ij}$. There should be found the location pattern \mathbf{x} lexicographically minimaximizing the vector of individual objective functions $(f_j(\mathbf{x}))_{j=1, \dots, m}$.

For the tests we used two-dimensional discrete location problems. The locations of the clients were defined by coordinates being the multiple of 5 generated as random numbers uniformly distributed in the interval $[0, 100]$. In the computations we used rectilinear distances. We tested solution times for different size

Table 1. Computation times (in seconds) for the ordered outcomes approach

number of clients (m)	number of facilities (p)						
	1	2	3	5	7	10	15
2	0.1						
5	0.0	0.0	0.1				
10	0.7	1.5	1.4	0.9	0.4		
15	4.7	15.2	14.6	10.8	6.5	4.4	
20	13.2	54.0	118.6	84.7	60.9	26.9	11.0
25	36.6	—	—	—	—	—	67.4

parameters m and p . All the experiments were performed on the PC computer with Pentium 4, 1.7 Ghz processor employing the CPLEX 9.1 package.

Tables 1 and 2 present solution times for two approaches being analyzed. The times are the averages of 10 randomly generated problems. The empty cell (minus sign) shows that the timeout of 120 seconds occurred. One can see fast growing times for the ordered outcomes approach with increasing number of clients (criteria). The growth is faster than the corresponding growth of the

Table 2. Computation times (in seconds) for the ordered values approach

number of clients (m)	number of facilities (p)						
	1	2	3	5	7	10	15
2	0.0						
5	0.1	0.0	0.0				
10	0.2	0.1	0.1	0.0	0.0		
15	0.7	0.7	0.3	0.1	0.1	0.0	
20	1.7	2.5	2.6	1.2	0.3	0.1	0.1
25	3.4	8.8	2.7	0.8	1.7	0.4	0.2

Table 3. Computation times (in seconds) of algorithms steps ($m = 20$)

step number	ordered outcomes approach			ordered values approach		
	$p = 1$	$p = 3$	$p = 5$	$p = 1$	$p = 3$	$p = 5$
	1	0.8	3.5	3.2	0.0	0.0
2	2.0	5.2	4.4	0.1	0.1	0.0
3	1.2	4.4	4.1	0.1	0.1	0.2
4	1.1	4.5	4.4	0.1	0.2	0.1
5	0.8	4.3	4.3	0.0	0.2	0.3
6	0.9	4.7	4.7	0.1	0.2	0.2
7	0.8	5.1	4.9	0.1	0.4	0.2
8	0.7	5.9	5.6	0.0	0.3	0.1
9	0.8	5.9	5.0	0.1	0.3	0.0
10	0.7	12.5	5.5	0.0	0.3	
11	0.9	7.7	6.5	0.1	0.3	
12	0.7	6.0	6.2	0.0	0.1	
13	0.4	6.8	5.7	0.1	0.1	
14	0.3	7.8	5.0	0.0	0.0	
15	0.3	8.1	4.4	0.0		
16	0.2	6.8	3.3	0.1		
17	0.2	6.6	2.8	0.0		
18	0.2	3.7	1.8	0.0		
19	0.1	6.3	1.5	0.1		
20	0.1	2.8	1.4	0.0		
21 – 30				0.6		
31 – 34				0.0		

problem sizes. Actually, it turns out that the solution times for the ordered outcomes model (5) are not significantly better (and in some instances even worse) than those for model (2), despite the latter uses auxiliary integer variables. On the other hand, the ordered values approach performs very well particularly with the number of the clients increasing. In fact, in this approach the number of steps depends not on the number of clients but on the number of different values of distances and this is constant. Moreover, the ordered values approach requires less steps for bigger number of facilities. This is due to the fact that the largest distance in the experiments does not exceed $200/p$.

Table 3 shows how introduction of the auxiliary constraints affects the performance in consecutive steps of the algorithms. One can notice that the ordered values technique generates MIP problems that are solved below 0.3 sec. (below 0.1 for $p = 1$) while the ordered outcomes problems require much longer computations. Despite a similar structure of auxiliary constraints in both approaches, the ordered values problems are much easier to solve. This property additionally contributes to the overall outperformance of the former and supports the attractiveness of the ordered values algorithm even for problems with the large number of different outcome values.

4 Concluding Remarks

The point-wise ordering of outcomes causes that the Lexicographic Min-Max optimization problem is, in general, hard to implement. We have analyzed optimization models allowing to form lexicographic sequential procedures for various nonconvex (possibly discrete) LMM optimization problems. Two approaches based on some LP expansion of the original model remain relatively simple for implementation independently of the problem structure. However, the ordered outcomes model performs similarly to the classical model with integer variables used to implement ordering and it is clearly outperformed by the ordered values approach. Further work on specialized algorithms (including heuristics) for the ordered values approach to various classes of discrete optimization problems seems to be a very promising research direction.

References

1. Behringer, F.A.: A simplex based algorithm for the lexicographically extended linear maxmin problem. *Eur. J. Opnl. Res.* **7** (1981) 274–283.
2. Bertsekas, D., Gallager, R.: *Data Networks*. Prentice-Hall, Englewood Cliffs (1987).
3. Burkard, R.E., Rendl, F.: Lexicographic bottleneck problems. *Oper. Res. Let.* **10** (1991) 303–308.
4. Della Croce, F., Paschos, V.T., Tsoukias, A.: An improved general procedure for lexicographic bottleneck problem. *Oper. Res. Let.* **24** (1999) 187–194.
5. Denda, R., Banchs, A., Effelsberg, W.: The fairness challenge in computer networks. *Lect. Notes Comp. Sci.* **1922** (2000) 208–220.
6. Dresher M.: *Games of Strategy*. Prentice-Hall, Englewood Cliffs (1961).

7. Dubois, D., Fortemps, Ph., Pirlot, M., Prade, H.: Leximin optimality and fuzzy set-theoretic operations. *Eur. J. Opnl. Res.* **130** (2001) 20–28.
8. Ehrgott, M.: Discrete decision problems, multiple criteria optimization classes and lexicographic max-ordering. *Trends in Multicriteria Decision Making*, T.J. Stewart, R.C. van den Honert (eds.), Springer, Berlin (1998), 31–44.
9. Klein, R.S., Luss, H., Smith, D.R.: A lexicographic minimax algorithm for multi-period resource allocation. *Math. Progr.* **55** (1992) 213–234.
10. Kostreva M.M., Ogryczak W. (1999) Linear optimization with multiple equitable criteria. *RAIRO Oper. Res.*, 33, 275–297.
11. Luss, H.: On equitable resource allocation problems: A lexicographic minimax approach. *Oper. Res.* **47** (1999) 361–378.
12. Marchi, E., Oviedo, J.A.: Lexicographic optimality in the multiple objective linear programming: the nucleolar solution. *Eur. J. Opnl. Res.* **57** (1992) 355–359.
13. Ogryczak, W.: *Linear and Discrete Optimization with Multiple Criteria: Preference Models and Applications to Decision Support* (in Polish). Warsaw Univ. Press, Warsaw (1997).
14. Ogryczak, W.: On the lexicographic minimax approach to location problems. *Eur. J. Opnl. Res.* **100** (1997) 566–585.
15. Ogryczak, W.: Multiple criteria optimization and decisions under risk. *Control & Cyber.* **31** (2002) 975–1003.
16. Ogryczak, W., Pióro, M., Tomaszewski, A.: Telecommunication network design and max-min optimization problem. *J. Telecom. Info. Tech.* **3/05** (2005) 43–56.
17. Ogryczak, W., Śliwiński, T.: On equitable approaches to resource allocation problems: the conditional minimax solution, *J. Telecom. Info. Tech.* **3/02** (2002) 40–48.
18. Ogryczak W., Śliwiński, T. (2003) On solving linear programs with the ordered weighted averaging objective. *Eur. J. Opnl. Res.*, 148, 80–91.
19. Ogryczak, W., Tamir, A.: Minimizing the sum of the k largest functions in linear time. *Inform. Proc. Let.* **85** (2003) 117–122.
20. Pióro, M., Medhi, D.: *Routing, Flow and Capacity Design in Communication and Computer Networks*. Morgan Kaufmann, San Francisco (2004).
21. Potters, J.A.M., Tijs, S.H.: The nucleolus of a matrix game and other nucleoli. *Math. of Oper. Res.* **17** (1992) 164–174.
22. Rawls, J.: *The Theory of Justice*. Harvard University Press, Cambridge (1971) .
23. Rice, J.R.: Tschebyscheff approximation in a compact metric space. *Bull. Amer. Math. Soc.* **68** (1962) 405–410.
24. Yager, R.R.: On the analytic representation of the Leximin ordering and its application to flexible constraint propagation. *Eur. J. Opnl. Res.* **102** (1997) 176–192.

Multivariate Convex Approximation and Least-Norm Convex Data-Smoothing

Alex Y.D. Siem¹, Dick den Hertog¹, and Aswin L. Hoffmann²

¹ Department of Econometrics and Operations Research/Center for Economic Research (CentER), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
`{a.y.d.siem, d.denhertog}@uvt.nl`

² Department of Radiation Oncology, Radboud University Nijmegen Medical Centre, Geert Grooteplein 32, 6525 GA Nijmegen, The Netherlands
`a.hoffmann@rther.umcn.nl`

Abstract. The main contents of this paper is two-fold. First, we present a method to approximate multivariate convex functions by piecewise linear upper and lower bounds. We consider a method that is based on function evaluations only. However, to use this method, the data have to be convex. Unfortunately, even if the underlying function is convex, this is not always the case due to (numerical) errors. Therefore, secondly, we present a multivariate data-smoothing method that smooths nonconvex data. We consider both the case that we have only function evaluations and the case that we also have derivative information. Furthermore, we show that our methods are polynomial time methods. We illustrate this methodology by applying it to some examples.

1 Introduction

In the field of discrete approximation, we are interested in approximating a function $y : \mathbb{R}^q \rightarrow \mathbb{R}$, given a discrete dataset $\{(x^i, y^i) : 1 \leq i \leq n\}$, where $x^i \in \mathbb{R}^q$ and $y^i = y(x^i) \in \mathbb{R}$, and n is the number of data points. It may happen that we know beforehand that the function $y(x)$ is convex. However, many approximation methods do not make use of the information that $y(x)$ is convex and construct approximations that do not preserve the convexity. For the univariate case there is some literature on convexity preserving functions; see e.g. [1] and [2]. In [1], Splines are used, and in [2], polynomial approximation is considered. For the multivariate case, in [3], convex quadratic polynomials are used to approximate convex functions. Furthermore, there is a lot of literature on so-called Sandwich algorithms; see e.g. [4], [5], [6], [7], and [8]. In these papers, upper and lower bounds for the function $y(x)$ are constructed, based on the discrete dataset, and based on the knowledge that $y(x)$ is convex.

A problem that may occur in practice is that one may have a dataset that is subject to noise, i.e., instead of the data y^i we have $\tilde{y}^i = y(x^i) + \varepsilon_y^i$, where ε_y^i is (numerical) noise. There may also be noise in the input data, i.e., $\tilde{x}^i = x^i + \varepsilon_x^i$, and if derivative information is available, it could also be subject to noise, i.e., $\tilde{\nabla}y^i = \nabla y^i + \varepsilon_g^i$, where $\nabla y^i = \nabla y(x^i)$. Note that we assume $y(x)$ to be convex. However, due to the noise, the perturbed data might loose the convexity of

$y(x)$, i.e., the noise could be such that it is not possible to fit a convex function through the perturbed data. Therefore, we are interested in data-smoothing, i.e., in shifting the data points, such that they obtain convexity, and such that the amount of movement of the data is minimized. This problem has already been tackled in literature for the univariate case; see e.g. [9], [10], and [11]. Also in isotonic regression, this problem is dealt with for the univariate case; see [12].

In this paper, we will consider two problems. First, we consider how to construct piecewise linear upper and lower bounds to approximate the output for the multivariate case. This extends the method in [7] to the multivariate case. If derivative information is available it is easy to construct upper and lower bounds. However, derivative information is not always available, e.g., in the case of black-box functions. In this paper, it turns out that these upper and lower bounds can be found by solving linear programs (LPs).

Second, we will consider the multivariate data-smoothing problem. We consider both the case that we have only function evaluations and the case that we also have derivative information. We will show that, if we only consider errors in the output data, the first problem can be solved by using techniques, which are from linear robust optimization; see [13]. It turns out that this problem can be tackled by solving an LP. If we also have derivative information, we can also consider errors in the gradients and in the input variables. We then obtain a nonlinear optimization problem. However, if we assume that there are only errors in the gradients and in the output data, we obtain an LP. Also, if we assume that there are only errors in the input data and in the output data, we also obtain an LP.

The remainder of this paper is organized as follows. In Sect. 2, we consider the problem of constructing upper and lower bounds. In Sect. 3, we consider multivariate data-smoothing, and in Sect. 4, we give some examples of the application of the data-smoothing techniques, considered in Sect. 3. Finally, in Sect. 5, we present possible directions for further research.

2 Bounds Preserving Convexity

In this section we assume that $y(x)$ is convex and that the data $(x^i, y(x^i))$ for $i = 1, \dots, n$ are convex as well, i.e., there are no (numerical) errors, and there exists a convex function that fits through the data points.

2.1 Upper Bounds

We are interested in finding the smallest upper bound for $y(x)$, given convexity, and the data $(x^i, y(x^i))$, for $i = 1, \dots, n$. Let $x = \sum_{i=1}^n \alpha_i x^i$, where $\sum_{i=1}^n \alpha_i = 1$, and $0 \leq \alpha_i \leq 1$, i.e., x is a convex combination of the input data x^i . Then, it is well-known that convexity gives us the following inequality:

$$y(x) = y\left(\sum_{i=1}^n \alpha_i x^i\right) \leq \sum_{i=1}^n \alpha_i y(x^i) .$$

This means that $\sum_{i=1}^n \alpha_i y(x^i)$ is an upper bound for $y(x)$. To find the smallest upper bound we should therefore solve

$$\begin{aligned}
 u(x) &:= \min_{\alpha_1, \dots, \alpha_n} \sum_{i=1}^n \alpha_i y(x^i) \\
 \text{s.t. } x &= \sum_{i=1}^n \alpha_i x^i \\
 0 &\leq \alpha_i \leq 1 \\
 \sum_{i=1}^n \alpha_i &= 1,
 \end{aligned} \tag{1}$$

where we put the decision variables underneath 'min'.

2.2 Lower Bounds

If we have derivative information, it is easy to construct a lower bound. It is well-known that if $y(x)$ is convex, we have that

$$y(x) \geq y(x^i) + \nabla y(x^i)^T (x - x^i), \quad \forall x \in \mathbb{R}^q, \forall i = 1, \dots, n .$$

Therefore, $\ell(x) = \max_{i=1, \dots, n} (y(x^i) + \nabla y(x^i)^T (x - x^i))$ is a lower bound.

If we do not have derivative information, we have to do something else. We are interested in finding the largest lower bound for $y(x)$, given convexity and the data $(x^i, y(x^i))$, for $i = 1, \dots, n$. Let $x^k = \sum_{i \neq k} \alpha_i^k x^i + \alpha^k x$, where $\sum_{i \neq k} \alpha_i^k + \alpha^k = 1$, with $0 \leq \alpha_i^k \leq 1$, and $0 < \alpha^k \leq 1$, for all $k = 1, \dots, n$, i.e., x^k is a convex combination of x^i , $i \neq k$, and x . Then the following holds for all $k \in \{1, \dots, n\}$:

$$y(x^k) = y \left(\sum_{i \neq k} \alpha_i^k x^i + \alpha^k x \right) \leq \sum_{i \neq k} \alpha_i^k y(x^i) + \alpha^k y(x) . \tag{2}$$

Without loss of generality we may assume that $\alpha^k > 0$. Then we can rewrite (2) as

$$y(x) \geq \frac{y(x^k) - \sum_{i \neq k} \alpha_i^k y(x^i)}{\alpha^k}, \text{ for } k = 1, \dots, n .$$

This inequality gives us a lower bound for $y(x)$. To obtain the largest lower bound we should solve the following problem:

$$\max_{k=1, \dots, n} \left\{ \begin{array}{l} \max_{\alpha^k, \alpha_i^k} \frac{y(x^k) - \sum_{i \neq k} \alpha_i^k y(x^i)}{\alpha^k} \\ \text{s.t. } x^k = \sum_{i \neq k} \alpha_i^k x^i + \alpha^k x \\ \sum_{i \neq k} \alpha_i^k + \alpha^k = 1 \\ 0 \leq \alpha_i^k \leq 1 \\ 0 < \alpha^k \leq 1 \end{array} \right\} . \tag{3}$$

This comes down to solving n nonlinear optimization problems, and taking the value of the largest solution. Note that the nonlinear optimization problems have

linear constraints and a fractional objective with linear numerator and denominator. These kinds of optimization problems can be rewritten into an LP; see [14].

This can be done as follows. Define $t^k := 1/\alpha^k$. We can now rewrite the inner optimization problem in (3) as

$$\begin{aligned} & \max_{\alpha^k, \alpha_i^k, t^k} t^k y(x^k) - \sum_{i \neq k} \alpha_i^k t^k y(x^i) \\ & \text{s.t. } x^k t^k = \sum_{i \neq k} \alpha_i^k t^k x^i + \alpha^k t^k x \\ & \quad \sum_{i \neq k} \alpha_i^k t^k + \alpha^k t^k = t^k \\ & \quad \alpha_i^k t^k \geq 0 \\ & \quad \alpha^k t^k = 1, \end{aligned}$$

where we multiplied all constraints by t^k . Now we define $z_i^k := \alpha_i^k t^k$ and $z^k := \alpha^k t^k$. We then get

$$\begin{aligned} & \max_{z^k, z_i^k, t^k} t^k y(x^k) - \sum_{i \neq k} z_i^k y(x^i) \\ & \text{s.t. } x^k t^k = \sum_{i \neq k} z_i^k x^i + z^k x \\ & \quad \sum_{i \neq k} z_i^k + z^k = t^k \\ & \quad z_i^k \geq 0 \\ & \quad z^k = 1. \end{aligned} \tag{4}$$

Note that (4) is an LP. Therefore, for the lower bound $\ell(x)$ we obtain the following:

$$\ell(x) := \max_{k=1, \dots, n} \left\{ \begin{array}{l} \max_{z^k, z_i^k, t^k} t^k y(x^k) - \sum_{i \neq k} z_i^k y(x^i) \\ \text{s.t. } x^k t^k = \sum_{i \neq k} z_i^k x^i + z^k x \\ \sum_{i \neq k} z_i^k + z^k = t^k \\ z_i^k \geq 0 \\ z^k = 1. \end{array} \right\}. \tag{5}$$

Note that the number of constraints in (5) is $q + 1$. The number of variables in (5) is also $q + 1$. Therefore it takes polynomial time to find the lower bound.

3 Convex Data-Smoothing

If the dataset is not convex, we first have to smooth the data such that it becomes convex. We distinguish between the case that we only have function evaluations and the case that we also have derivative information.

3.1 Function Value Information

We only consider movement of the output data \tilde{y}^i . So, we want to minimally shift the perturbed output data \tilde{y}^i such that they become convex. In the following

$$\begin{cases} (x^i)^T r^i + v^i \geq y_s^i \\ (x^k)^T r^i + v^i \leq y_s^k \quad \forall k \neq i \\ r^i \in \mathbb{R}^q, v^i \in \mathbb{R} \end{cases} \quad (10)$$

We can now finally rewrite the second constraint in (6) as (10) for every $i = 1, \dots, n$. This means that we can rewrite (6) as

$$\begin{aligned} \min_{\delta_y^+, \delta_y^-, y_s, r^i, v^i} & \sum_{i=1}^n ((\delta_y^+)^i + (\delta_y^-)^i) \\ \text{s.t.} & y_s^i = \tilde{y}^i + (\delta_y^+)^i - (\delta_y^-)^i \quad \forall i = 1, \dots, n \\ & (x^i)^T r^i + v^i \geq y_s^i \quad \forall i = 1, \dots, n \\ & (x^k)^T r^i + v^i \leq y_s^k \quad \forall k \neq i, \forall i = 1, \dots, n \\ & \delta_y^+ \in \mathbb{R}_+^n, \delta_y^- \in \mathbb{R}_+^n, \\ & r^i \in \mathbb{R}^q, v^i \in \mathbb{R} \quad \forall i = 1, \dots, n, \end{aligned} \quad (11)$$

which is an LP. Note that, after substituting the equality constraints for y_s^i , the number of constraints in (11) is $n(n - 1) + n = n^2$. The number of variables in (11) is $(q + 3)n$.

Above, we minimized the sum of the absolute values of the shifts, i.e. the ℓ_1 -norm. However, we can also choose to minimize other norms, such as e.g., the ℓ_∞ -norm or the ℓ_2 -norm. Using the ℓ_∞ -norm, we also obtain an LP, which is similar to (11).

3.2 Derivative Information

Next, we consider the case in which we also have gradient information. Suppose that the underlying function is convex, but the data are not convex, due to (numerical) errors. Again, we are interested in shifting the data such that they become convex. We consider perturbed output values \tilde{y}^i , perturbed gradients $\tilde{\nabla}y(x^i)$, and perturbed input values \tilde{x}^i . Therefore in this case we want to minimize the shifts in the output values, in the gradients, and in the inputs. So, in the following optimization problem, we minimize the sum of upward and downward shifts $(\delta_y^+)^i$ and $(\delta_y^-)^i$ of the output values, the upward and downward shifts $(\delta_g^+)^i$ and $(\delta_g^-)^i$ of the gradient, and the upward and downward shifts $(\delta_x^+)^i$ and $(\delta_x^-)^i$ of the input values such that the data become convex:

$$\begin{aligned} \min_{\substack{(\delta_y^+)^i, (\delta_y^-)^i, (\delta_g^+)^i, \\ (\delta_g^-)^i, (\delta_x^+)^i, (\delta_x^-)^i, \\ x_s^i, y_s^i, (\nabla y^i)_s}} & \sum_{i=1}^n ((\delta_y^+)^i + (\delta_y^-)^i + e_q^T (\delta_g^+)^i + e_q^T (\delta_g^-)^i + e_q^T (\delta_x^+)^i + e_q^T (\delta_x^-)^i) \\ \text{s.t.} & (\nabla y^i)_s = \tilde{\nabla}y^i + (\delta_g^+)^i - (\delta_g^-)^i \quad \forall i = 1, \dots, n \\ & x_s^i = \tilde{x}^i + (\delta_x^+)^i - (\delta_x^-)^i \quad \forall i = 1, \dots, n \\ & y_s^i = \tilde{y}^i + (\delta_y^+)^i - (\delta_y^-)^i \quad \forall i = 1, \dots, n \\ & (\nabla y^i)_s^T (x_s^j - x_s^i) + y_s^i \leq y_s^j \quad \forall i, j = 1, \dots, n, i \neq j \\ & (\delta_y^+)^i \in \mathbb{R}_+, (\delta_y^-)^i \in \mathbb{R}_+, (\delta_g^+)^i \in \mathbb{R}_+^q \quad \forall i = 1, \dots, n \\ & (\delta_g^-)^i \in \mathbb{R}_+^q, (\delta_x^+)^i \in \mathbb{R}_+^q, (\delta_x^-)^i \in \mathbb{R}_+^q \quad \forall i = 1, \dots, n, \end{aligned} \quad (12)$$

where $\nabla y^i = \nabla y(x^i)$, and e_q is the q -dimensional all-one vector. The 4-th constraint in (12) is a necessary and sufficient condition for convexity of the data; see page 338 in [15]. However, (12) is a nonconvex optimization problem, and therefore not tractable.

However, if there is no uncertainty in the input values x^1, \dots, x^n , we can omit the variables $(\delta_x^+)^i$ and $(\delta_x^-)^i$ in (12), and we obtain an LP. Similarly, if there is no uncertainty in the values of the gradients, we can omit $(\delta_g^+)^i$ and $(\delta_g^-)^i$ in (12), and we also obtain an LP.

An example of a problem, where the gradient information is exact, and we only have errors in the input variables and output variables is in the field of multiobjective optimization. In the so-called weighted sum method, to determine a point on the Pareto curve/surface the weights determine the exact value of the gradient, whereas due to numerical errors of the solver, the input value and the output value might be subject to noise.

Note that in the formulation of (12) we have minimized the shifts $(\delta_y^+)^i, (\delta_y^-)^i, (\delta_g^+)^i, (\delta_g^-)^i, (\delta_x^+)^i, (\delta_x^-)^i$, and have given them all equal importance. However, we might want to give one type of the error more weight than the other type.

4 Numerical Examples

In this section we will consider some examples of the theory discussed in Sect. 3.

Example 1 (artificial, no derivative information). In this example we apply the theory that we developed in Sect. 3.1. We consider the function $y : \mathbb{R}^2 \rightarrow \mathbb{R}$, given by $y(x_1, x_2) = x_1^2 + x_2^2$. We take a sample of 10 input data points x^1, \dots, x^{10} from $[-2, 2] \times [-2, 2]$, and calculate the output values $y(x^1), \dots, y(x^{10})$. Furthermore, we add some noise to it, i.e., we add a noise ε_y^i , where the ε_y^i 's are independent and uniformly distributed on $[-2.5, 2.5]$, such that the data become nonconvex. We obtain values $\tilde{y}^i = y^i + \varepsilon_y^i$. The values are given in Table 1. We solved (11) for this problem, and the shifted data y_s^i are also given in Table 1. The values that are really shifted, are shown in italics. □

Table 1. Data and results of smoothing in Example 1

number	x_1	x_2	y	\tilde{y}	y_s
1	-0.0199	-1.9768	3.9081	6.1588	6.1588
2	0.0925	1.3411	1.8071	0.4628	0.4628
3	1.4427	0.3253	2.1872	2.7214	2.7214
4	-1.8056	-1.1961	4.6908	4.6208	4.6208
5	-0.4435	-0.3444	0.3153	2.2718	<i>2.0578</i>
6	-1.2952	0.8811	2.4539	3.7644	3.7644
7	1.7826	1.6795	5.9984	5.7807	5.7807
8	0.8074	-1.3585	2.4974	0.0899	<i>0.4842</i>
9	-0.8714	0.5089	1.0183	2.6254	2.6254
10	0.5779	-0.7205	0.8531	0.5766	0.5766

Example 2 (radio therapy, no derivative information). In radiotherapy the main goal is to give the tumour enough radiation dose, such that the surrounding organs do not receive too much radiation dose. This problem can be formulated mathematically by a multiobjective optimization problem. With the tumour and each healthy surrounding organ, an objective function is associated. One of the problems is that the calculation of a point on the Pareto surface can be very time-consuming. Therefore, we are interested in approximating the Pareto surface; see e.g. [16]. Under certain conditions, we may assume that this Pareto surface is convex. However, due to numerical errors, the Pareto points may not be convex. Therefore we should first smooth them to make them convex.

We have data from a patient of the Radboud University Nijmegen Medical Centre, in Nijmegen, the Netherlands. This data is from a multiobjective optimization problem with 3 objectives, and has 69 data points. The data are shown in Fig. 1. The Pareto surface is a convex and decreasing function. However, it turned out that the data is not convex. By solving (11), the data is smoothed such that the data becomes convex. The smoothed data points are also shown in Fig. 1. □

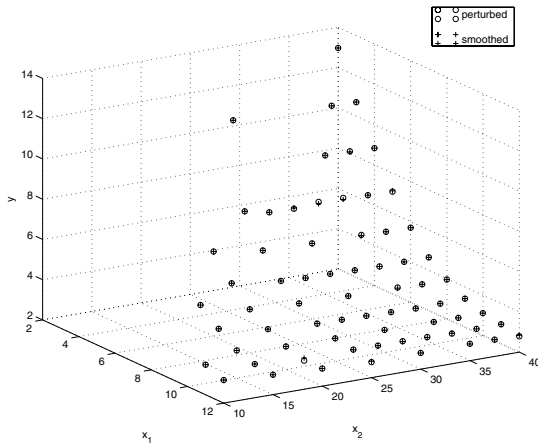


Fig. 1. The the perturbed data and the smoothed data of Example 2

5 Further Research

As interesting topics for further research we mention several possible applications of the methods developed in this paper.

Possible applications for the construction of the bounds in Sect. 2 are:

- Extend the Sandwich algorithms as exist for the approximation of univariate convex functions to the multivariate case by using the lower and upper bounds. More specifically, this may be useful for approximating convex Pareto surfaces and black-box functions (e.g. deterministic computer simulation).

- Use the lower bounds in convex optimization. For each new candidate proposed by the nonlinear programming solver, we can calculate the lower bound, and if this lower bound is larger than the best known objective value up to now, we reject the candidate before evaluating its function value. This may reduce computation time, especially when the function evaluation is time-consuming. In [17] promising results are shown for the univariate case.

Possible applications for the data-smoothing methods of Sect. 3 are:

- Apply data-smoothing before applying Sandwich type algorithms. This may be necessary because of (numerical) noise. This noise occurs e.g., when we want to estimate a Pareto surface in the field of multiobjective optimization.
- Apply data-smoothing to multivariate isotonic regression problems.
- Apply the data-smoothing techniques to sampling methods to assess the convexity/concavity of multivariate nonlinear functions; see [18].

References

1. Kuijt, F.: Convexity preserving interpolation – Stationary nonlinear subdivision and splines. PhD thesis, University of Twente, Enschede, The Netherlands (1998)
2. Siem, A.Y.D., de Klerk, E., den Hertog, D.: Discrete least-norm approximation by nonnegative (trigonometric) polynomials and rational functions. CentER Discussion Paper 2005-73, Tilburg University, Tilburg (2005)
3. den Hertog, D., de Klerk, E., Roos, K.: On convex quadratic approximation. *Statistica Neerlandica* **563** (2002) 376–385
4. Burkard, R.E., Hamacher, H.W., Rote, G.: Sandwich approximation of univariate convex functions with an application to separable convex programming. *Naval Research Logistics* **38** (1991) 911–924
5. Fruhwirth, B., Burkard, R.E., Rote, G.: Approximation of convex curves with application to the bi-criteria minimum cost flow problem. *European Journal of Operational Research* **42** (1989) 326–338
6. Rote, G.: The convergence rate of the Sandwich algorithm for approximating convex functions. *Computing* **48** (1992) 337–361
7. Siem, A.Y.D., den Hertog, D., Hoffmann, A.L.: A method for approximating univariate convex functions using only function evaluations. Working paper, Tilburg University, Tilburg (2005)
8. Yang, X.Q., Goh, C.J.: A method for convex curve approximation. *European Journal of Operational Research* **97** (1997) 205–212
9. Cullinan, M.P.: Data smoothing using non-negative divided differences and ℓ_2 approximation. *IMA Journal of Numerical Analysis* **10** (1990) 583–608
10. Demetriou, I.C., Powell, M.J.D.: Least squares smoothing of univariate data to achieve piecewise monotonicity. *IMA Journal of Numerical Analysis* **11** (1991) 411–432
11. Demetriou, I.C., Powell, M.J.D.: The minimum sum of squares change to univariate data that gives convexity. *IMA Journal of Numerical Analysis* **11** (1991) 433–448
12. Barlow, R.E., Bartholomew, R.J., Bremner, J.M., Brunk, H.D.: *Statistical inference under order restrictions*. Wiley, Chichester (1972)
13. Ben-Tal, A., Nemirovski, A.: *Robust optimization - methodology and applications*. *Mathematical Programming, Series B* **92** (2002) 453–480

14. Charnes, A., Cooper, W.W.: Programming with linear fractional functional. *Naval Research Logistics Quarterly* **9** (1962) 181–186
15. Boyd, S., Vandenberghe, L.: *Convex optimization*. Cambridge University Press, Cambridge (2004)
16. Hoffmann, A.L., Siem, A.Y.D., den Hertog, D., Kaanders, J.H.A.M., Huizenga, H.: Dynamic generation and interpolation of pareto optimal IMRT treatment plans for convex objective functions. Working paper, Radboud University Nijmegen Medical Centre, Nijmegen (2005)
17. den Boef, E., den Hertog, D.: Efficient line searching for convex functions. CentER Discussion Paper 2004-52, Tilburg University, Tilburg (2004)
18. Chinneck, J.W.: Discovering the characteristics of mathematical programs via sampling. *Optimization Methods and Software* **17**(2) (2002) 319–352

Linear Convergence of Tatônnement in a Bertrand Oligopoly*

Guillermo Gallego¹, Woonghee Tim Huh¹, Wanmo Kang^{1,**},
and Robert Phillips²

¹ Columbia University, New York, USA

² Stanford University and Nomis Solutions, USA

Abstract. We show the linear convergence of the tatônnement scheme in a Bertrand oligopoly price competition game using a possibly asymmetric attraction demand model with convex costs. To demonstrate this, we also show the existence of the equilibrium.

1 Introduction and Model

In the Bertrand oligopoly price competition model for differentiated products, a variety of demand models and cost models have been used. The choice of these models affects the profit of each firm.

We let n be the number of firms, which are indexed by $i = 1, \dots, n$. The demand for each firm is specified as a function of prices. Let p_i denote the price of firm i , and define the price vector of competing firms by \mathbf{p}_{-i} , which is a shorthand for $(p_1, \dots, p_{i-1}, p_{i+1}, \dots, p_n)$. Also denote the vector of all prices by $\mathbf{p} = (p_1, \dots, p_n) = (p_i, \mathbf{p}_{-i})$. The demand for each firm i is given by $d_i = d_i(\mathbf{p})$. We assume that firm i 's demand is strictly decreasing in its price (i.e., $\partial d_i / \partial p_i < 0$), and that products are gross substitutes (i.e., $\partial d_i / \partial p_j \geq 0$ whenever $j \neq i$).

In this paper we consider a generalization of the logit demand model called the *attraction demand model*:

$$d_i(\mathbf{p}) := \frac{a_i(p_i)}{\sum_j a_j(p_j) + \kappa} \quad (1)$$

where κ is either 0 or strictly positive. As discussed in [2] and [14], the attraction model has successfully been used in estimating demand in econometric studies, and is increasingly accepted in marketing, e.g., [5]. For its applications in the operations management community, see [21], [4] and the references therein. Without any loss of generality, we normalize the total demand of the market to 1. The *attraction function* $a_i(\cdot)$ of firm i is a positive and strictly decreasing function of its price.

For the cost model, we assume that cost is not a function of price, but of demand alone. We denote firm i 's cost function by $C_i(d_i)$ defined on $[0, 1]$ and

* For the full version of this paper, see [11].

** Corresponding author.

assume C_i is increasing and convex. The profit of firm i is the difference between its revenue and cost, given by

$$\pi_i := \pi_i(\mathbf{p}) := p_i \cdot d_i(\mathbf{p}) - C_i(d_i(\mathbf{p})). \tag{2}$$

Each firm’s objective is to maximize π_i . We impose technical conditions on the attraction demand and cost models as outlined in Section 2.

The study of oligopolistic interaction is a classical problem in economics. In the model proposed by Cournot, firms compete on production output quantities, which in turn determine the market price. In Bertrand’s model, however, competition is based on prices instead of production quantities. In the price competition models by Edgeworth, each firm decides how much of its demand is satisfied, in which case an equilibrium solution may or may not exist. Price competition with product differentiation has also been studied by [12], [20] and [8]. An extensive treatment of the subject is found in [24]. We provide a brief summary of results regarding the existence, and stability of equilibrium, followed by their application in the operations management literature.

Existence. There are two common methods to show existence of an equilibrium in price competition games. The first method is to obtain existence through the quasi-concavity of π_i in p_i . See [7]. The second method shows existence through *supermodular* games. See [22] for the existence of an equilibrium in supermodular games, and [16], for monotone transformation of supermodular games. Thus, if the price competition game is supermodular, it has at least one equilibrium. Similarly, [17] shows the existence of a Nash equilibrium for a generalization of supermodular games, called games with strategic complementarities. Such games include instances of price competition. These however are not applicable to our model.

Stability. By definition, a set of actions at equilibrium is a fixed point of the best response mapping. A simultaneous discrete *tatônnement* is a sequence of actions in which the current action of each firm is the best response to the previous actions of other firms. An equilibrium is *globally stable* if the tatônnement converges to this equilibrium regardless of the initial set of actions. [23] shows that if a supermodular game with continuous payoffs has a unique equilibrium, it is globally stable. To our knowledge, there is no known result regarding the provable convergence rate of the tatônnement in the price competition game.

Operations Management Literature. There has been a growing interest in oligopolistic price competition in the operations literature. To predict and study market outcomes, the existence and the uniqueness of equilibrium are often required. Stability and convergence rate indicate both the robustness of equilibrium and the efficiency of computational algorithms. For example, see [4], [3], [1], and [6].

2 Assumptions

This section lists our assumptions on the attraction function $a_i(\cdot)$ in (1), and the profit function π_i and the cost function $C_i(\cdot)$ in (2).

We let $\rho_i := \inf\{p : a_i(p) = 0\}$ be the upper bound on price p_i . Firm i 's action space for price is an open interval $(0, \rho_i)$. Let $\mathcal{P} := (0, \rho_1) \times \dots \times (0, \rho_n)$. Let $\eta_i(p) := -p \cdot a'_i(p) / a_i(p)$ be the *elasticity* of firm i 's attraction function. We adopt the following simplifying notation: $f(x+) := \lim_{h \downarrow x} f(h)$, $f(x-) := \lim_{h \uparrow x} f(h)$, $\inf \emptyset = \infty$, and $\frac{y}{y+k} = 1$ if $y = \infty$ and k is finite.

Condition A: (A1) $a_i(\cdot)$ is positive, strictly decreasing and continuously differentiable, i.e., $a_i(p) > 0$ and $a'_i(p) < 0$ for all $p \in (0, \rho_i)$.

(A2) The elasticity $\eta_i(\cdot)$ is nondecreasing.

(A3) If $a_i(0+) < \infty$, then $a'_i(0+) > -\infty$.

Condition B: (B1) $C_i(\cdot)$ is strictly increasing, continuously differentiable, and convex on $[0, 1]$, i.e., $c_i(\cdot) := C'_i(\cdot)$ is positive and increasing, and satisfies $c_i(0+) > 0$.

Condition C: (C1) $c_i(0) < \rho_i \cdot (1 - 1/\eta_i(\rho_i))$, i.e., the Lerner index $[p_i - c_i(d_i)]/p_i$ at price $p_i = \rho_i$ and demand $d_i = 0$ is strictly bounded below by $1/\eta_i(\rho_i)$.

(C2) If $\kappa = 0$, then $c_i(1) < \rho_i$.

(C3) If $\kappa = 0$, the following technical condition holds:

$$\sum_{i=1}^n \left(1 - \frac{1}{\eta_i(\rho_i) \cdot (1 - c_i(1)/\rho_i)} \right) > 1 .$$

Examples of $C_i(\cdot)$ include the linear function and exponential function. More examples are provided in Section 3. We remark that attraction functions do not need to be identical. Furthermore, even the form of the attraction function may not be same among firms. Analogously, the cost functions need not have the same form.

For the rest of this paper, we assume Conditions A, B and C hold. In Section 5, we introduce an additional assumption that both $C_i(\cdot)$ and $a_i(\cdot)$ are twice continuously differentiable.

3 Examples

In this section, we list price competition models in which the convex cost model is applicable. We present some examples from an inventory-capacity systems and a service system based on queues.

Inventory-Capacity System: In the first example, consider the pricing problem in the stochastic inventory system with exogenously determined stocking levels. We denote stochastic demand of firm i by $D_i(\mathbf{p})$, and its expected demand by $d_i(\mathbf{p})$. Demand is a function of the price vector $\mathbf{p} = (p_1, \dots, p_n)$. We represent firm i 's stochastic demand by $D_i(p_1, \dots, p_n) = \varphi(d_i(\mathbf{p}), \varepsilon_i)$, where ε_i is a random variable. (We can allow φ to be dependent on i). We suppose the continuous density function $f_i(\cdot)$ for ε_i exists, and let $F_i(\cdot)$ denote its cumulative density function.

Let y_i be the exogenously fixed stocking level of firm i . For the first y_i units, the per-unit materials cost is w_i . If realized demand is at most y_i , the per-unit

salvage value of $w_i - h_i > 0$ is obtained. Otherwise, the excess demand is met through an emergency supply at the cost of $w_i + b_i$ per unit, where $b_i \geq 0$. The profit of firm i is the difference between its revenue and costs, and the expected profit is $\pi_i(\mathbf{p}|y_i) = p_i \cdot d_i(\mathbf{p}) - C_i(d_i(\mathbf{p}), y_i)$, where

$$C_i(d_i, y_i) = w_i d_i + h_i E[y_i - \varphi(d_i, \varepsilon_i)]^+ + b_i E[\varphi(d_i, \varepsilon_i) - y_i]^+,$$

and h_i and b_i are the per-unit inventory overage and underage costs, respectively.

Our goal is to show that for fixed y_i , this function satisfies condition (B1). We achieve this goal with two common demand uncertainty models.

- Additive Demand Uncertainty Model: $\varphi(d_i, \varepsilon_i) = d_i + \varepsilon_i$ where $E[\varepsilon_i] = 0$. Then,

$$\begin{aligned} \frac{\partial C_i(d_i, y_i)}{\partial d_i} &= w_i - h_i P[y_i \geq d_i + \varepsilon_i] + b_i P[y_i \leq d_i + \varepsilon_i] \\ &= w_i - h_i F_i(y_i - d_i) + b_i(1 - F_i(y_i - d_i)). \end{aligned}$$

- Multiplicative Demand Uncertainty Model: $\varphi(d_i, \varepsilon_i) = d_i \cdot \varepsilon_i$ where ε_i is positive and $E[\varepsilon_i] = 1$. Then,

$$\begin{aligned} \frac{\partial C_i(d_i, y_i)}{\partial d_i} &= w_i - h_i \int_0^{y_i/d_i} \varepsilon dF_i(\varepsilon) + b_i \int_{y_i/d_i}^\infty \varepsilon dF_i(\varepsilon) \\ &= w_i - h_i + (h_i + b_i) \int_{y_i/d_i}^\infty \varepsilon dF_i(\varepsilon). \end{aligned}$$

In both cases, $\partial C_i(d_i, y_i)/\partial d_i$ is positive since $w_i > h_i$ and nondecreasing in d_i . We conclude that for fixed y_i , $C_i(d_i, y_i)$ is strictly increasing, twice continuously differentiable, and convex in d_i . Furthermore, $\partial C_i(d_i, y_i)/\partial d_i > 0$ at $d_i = 0$.

Service System: In the second example, we model each firm as a single server queue with finite buffer, where the firms' buffer sizes are given exogenously. Let κ_i denote the size of firm i 's buffer; no more than κ_i customers are allowed to the system. We assume exponential service times and the Poisson arrival process. The rate μ_i of service times are exogenously determined, and the rate d_i of Poisson arrival is an output of the price competition. In the queueing theory notation, each firm i is a $M/M/1/\kappa_i$ system.

We assume that the materials cost is $w_i > 0$ per served customer, and the diverted customers' demand due to buffer overflow is met by an emergence supply at the cost of $w_i + b_i$ unit per customer, where $b_i > 0$. The demand arrival rate $d_i = d_i(\mathbf{p})$ is determined as a function of the price vector \mathbf{p} . It follows that firm i 's time-average revenue is $p_i \cdot d_i - C_i(d_i)$, where $C_i(d_i)$ is the sum of $w_i \cdot d_i$ and the time-average number of customers diverted from the system is multiplied by b_i . Thus, according to elementary queueing theory (see, for example, [15]),

$$\begin{aligned} C_i(d_i) &= w_i \cdot d_i + b_i \cdot \frac{d_i \cdot (1 - d_i/\mu_i)(d_i/\mu_i)^{\kappa_i}}{1 - (d_i/\mu_i)^{\kappa_i+1}}, & \text{if } d_i \neq \mu_i \\ &= w_i \cdot d_i + b_i \cdot \frac{d_i}{\kappa_i + 1}, & \text{if } d_i = \mu_i. \end{aligned}$$

Algebraic manipulation shows that $C_i(\cdot)$ is convex and continuously twice differentiable, satisfying $c_i(0) = w_i > 0$.

4 Existence of Equilibrium

In this section, we show that the oligopoly price competition has a unique equilibrium. Given the price vector, each firm’s profit function is given by expression (2), where its demand is determined by (1). We first show that the first order condition $\partial\pi_i/\partial p_i = 0$ is sufficient for the Nash equilibrium (Proposition 1). For each value of a suitably defined aggregate attraction δ , we show that there is at most one candidate for the solution of the first order condition (Proposition 2). Then, we demonstrate that there exists a unique value δ of the aggregator such that this candidate indeed solves the first order condition (Propositions 3 and 4). We proceed by assuming both Conditions A, B and C. Let $c_i(\mathbf{p}) := \eta_i(p_i) \cdot (1 - d_i(\mathbf{p}))$.

Proposition 1. *Firm i ’s profit function π_i is strictly quasi-concave in $p_i \in (0, \rho_i)$. If $\mathbf{p}^* = (p_1^*, \dots, p_n^*) \in \mathcal{P}$ satisfies $\partial\pi_i(\mathbf{p}^*)/\partial p_i = 0$ for all i , \mathbf{p}^* is the Nash equilibrium in \mathcal{P} , and $p_i^* > c_i(0)$ for each i . Furthermore, the condition $\partial\pi_i/\partial p_i = 0$ is equivalent to*

$$c_i(d_i(\mathbf{p}))/p_i = 1 - 1/c_i(\mathbf{p}) . \tag{3}$$

Given a price vector, let $\delta := \sum_{j=1}^n a_j(p_j)$ be the aggregate attraction. The support of δ is $\Delta := (0, \sum_{j=1}^n a_j(0+))$. From (A1), it follows that $\delta \in \Delta$. Then, $d_i = a_i(p_i)/(\delta + \kappa)$. Since a_i^{-1} is well-defined by (A1), we get $p_i = a_i^{-1}((\delta + \kappa)d_i)$. Thus, (3) is equivalent to

$$\frac{c_i(d_i)}{a_i^{-1}((\delta + \kappa)d_i)} = 1 - \frac{1}{\eta_i \circ a_i^{-1}((\delta + \kappa)d_i) \cdot (1 - d_i)} . \tag{4}$$

Observe that there is one-to-one correspondence between $\mathbf{p} = (p_1, \dots, p_n)$ and $\mathbf{d} = (d_1, \dots, d_n)$, given δ (and of course, κ). Let $D_i(\delta)$ be the solution to (4) given δ (and κ). The existence and uniqueness of $D_i(\delta)$ are guaranteed by Proposition 2 below. The $D_i(\delta)$ ’s may not sum up to the “correct” value of $\delta/(\delta + \kappa)$ unless a set of conditions is satisfied (Propositions 3). Proposition 4 shows the existence of a unique δ such that the $D_i(\delta)$ ’s sum up to $\delta/(\delta + \kappa)$.

Let $\bar{d}_i(\delta) := \min \left\{ \frac{a_i(0+)}{\delta + \kappa}, 1 \right\}$ be an upper bound on the market share of firm i . For each fixed $\delta \in \Delta$, we define the following real-valued functions on $(0, \bar{d}_i(\delta))$:

$$L_i(x_i|\delta) := \frac{c_i(x_i)}{a_i^{-1}((\delta + \kappa)x_i)} \text{ and } R_i(x_i|\delta) := 1 - \frac{1}{\eta_i \circ a_i^{-1}((\delta + \kappa)x_i) (1 - x_i)} .$$

We remark that both $L_i(x_i|\delta)$ and $R_i(x_i|\delta)$ are continuous in x_i in $(0, \bar{d}_i(\delta))$.

Proposition 2. *For each i and each $\delta \in \Delta$, $L_i(\cdot|\delta)$ is positive and strictly increasing, and $R_i(\cdot|\delta)$ is strictly decreasing. Furthermore, $L_i(x_i|\delta) = R_i(x_i|\delta)$ has a unique solution in $(0, \bar{d}_i(\delta))$, i.e., $D_i(\delta)$ is a well-defined function of δ .*

For any aggregate attraction $\delta \in \Delta$, Proposition 2 shows that there is a unique solution x_i satisfying $L_i(x_i|\delta) = R_i(x_i|\delta)$, and this solution is $D_i(\delta)$. It represents demand that maximizes firm i 's profit provided that the aggregate attraction remains at δ . Also, define $D(\delta) := D_1(\delta) + \dots + D_n(\delta)$. Note that $D_i(\delta)$ is a strictly decreasing function since any increase in δ lifts the graph of $L_i(x_i|\delta)$ and drops that of $R_i(x_i|\delta)$. Therefore $D(\delta)$ is also a strictly decreasing function. Furthermore, $D_i(\delta)$ is a continuous function of δ . Thus, $D(\delta)$ is continuous.

Proposition 3. *For fixed $\delta \in \Delta$, $D(\delta) = \frac{\delta}{\delta + \kappa}$ holds if and only if there exist $\mathbf{p} = (p_1, \dots, p_n)$ and $\mathbf{d} = (d_1, \dots, d_n)$ such that the following set of conditions hold: (i) $\delta = \sum_{j=1}^n a_j(p_j)$, (ii) $d_i = a_i(p_i)/(\delta + \kappa)$ for each i , and (iii) $L_i(d_i|\delta) = R_i(d_i|\delta)$ for each i . In such case, furthermore, the price vector corresponding to any δ satisfying $D(\delta) = \frac{\delta}{\delta + \kappa}$ is unique.*

If there is $\delta \in \Delta$ satisfying $D(\delta) = \frac{\delta}{\delta + \kappa}$, then by Proposition 3, the corresponding price vector satisfies $\partial \pi_i / \partial p_i = 0$ for all i . By Proposition 1, this price vector is a Nash equilibrium. For the unique existence of the equilibrium, it suffices to show the result of the following proposition.

Proposition 4. *There exists a unique $\delta \in \Delta$ such that $D(\delta) = \delta / (\delta + \kappa)$.*

Theorem 1. *There exists a unique positive pure strategy Nash equilibrium price vector $\mathbf{p}^* \in \mathcal{P}$. Furthermore, \mathbf{p}^* satisfies $p_i^* > c_i(0)$ for all $i = 1, \dots, n$.*

5 Convergence of Tatônnement Scheme

In this section, we show that the unique equilibrium is globally stable under the tatônnement scheme. Suppose each firm i chooses a best-response pricing strategy: choose p_i maximizing his profit $\pi_i(p_1, \dots, p_n)$ while p_j 's are fixed for all $j \neq i$.

By Theorem 1, there exists a unique equilibrium vector, which is denoted by $\mathbf{p}^* = (p_1^*, \dots, p_n^*) \in \mathcal{P}$. Define $\mathcal{Q} := (0, a_1(0+)) \times \dots \times (0, a_n(0+))$. Let $\mathbf{q}^* = (q_1^*, \dots, q_n^*) \in \mathcal{Q}$ be the corresponding attraction vector where $q_i^* := a_i(p_i^*)$. Let $\hat{q}_i := \sum_{j \neq i} q_j$ be the sum of attraction quantities of firms other than i . Set $\theta_i^* := q_i^* / (\hat{q}_i^* + \kappa)$ and $d_i^* := q_i^* / (q_i^* + \hat{q}_i^* + \kappa)$, which are both positive. Suppose we fix the price p_j for all $j \neq i$, and let $q_i := a_i(p_i)$ be the corresponding attraction. Since a_i is one-to-one and $\delta = q_i + \hat{q}_i$, condition (4) is equivalent to

$$\frac{c_i\left(\frac{q_i}{q_i + \hat{q}_i + \kappa}\right)}{a_i^{-1}(q_i)} = 1 - \frac{1}{\eta_i \circ a_i^{-1}(q_i)} \left(1 + \frac{q_i}{\hat{q}_i + \kappa}\right). \tag{5}$$

Using an argument similar to Proposition 2 and ensuing discussion, it can be shown that there is a unique solution q_i to (5) for each \hat{q}_i given by any positive number less than $\sum_{j \neq i} a_i(0+)$. We call this solution q_i the *best response function* $\psi_i(\hat{q}_i)$ for firm i . The unique equilibrium satisfies $\psi_i(\hat{q}_i^*) = q_i^*$ where $\hat{q}_i^* = \sum_{j \neq i} q_j^*$. Furthermore, it is easy to show that $\psi_i(\cdot)$ is strictly increasing.

Proposition 5. $\psi_i(\cdot)$ is a strictly increasing function.

From the definition of θ_i^* and $\psi_i(\hat{q}_i^*) = q_i^*$, we know $\psi_i(\hat{q}_i)/(\hat{q}_i + \kappa) = \theta_i^*$ at $\hat{q}_i = \hat{q}_i^*$. The following proposition characterizes the relationship between $\psi_i(\hat{q}_i)/(\hat{q}_i + \kappa)$ and θ_i^* .

Proposition 6. $\psi_i(\hat{q}_i)/(\hat{q}_i + \kappa)$ is strictly decreasing in \hat{q}_i , and satisfies $\psi_i(\hat{q}_i^*)/(\hat{q}_i^* + \kappa) = \theta_i^*$. Thus,

$$\frac{\psi_i(\hat{q}_i)}{\hat{q}_i + \kappa} \begin{cases} > \theta_i^*, \text{ for } \hat{q}_i < \hat{q}_i^* \\ = \theta_i^*, \text{ for } \hat{q}_i = \hat{q}_i^* \\ < \theta_i^*, \text{ for } \hat{q}_i > \hat{q}_i^*. \end{cases}$$

Furthermore, $\psi_i'(\hat{q}_i)$ is continuous, and satisfies $0 < \psi_i'(\hat{q}_i^*) < \theta_i^*$.

Let $\mathbf{q} = (q_1, \dots, q_n)$. We denote the vector of best response functions by $\Psi(\mathbf{q}) = (\psi_1(\hat{q}_1), \dots, \psi_n(\hat{q}_n)) \in \mathcal{Q}$, where $\hat{q}_i = \sum_{j \neq i} q_j$. Note that $\mathbf{q}^* = (q_1^*, \dots, q_n^*)$ is a fixed point of Ψ , i.e., $\Psi(\mathbf{q}^*) = \mathbf{q}^*$. By Proposition 5, we have $\Psi(\mathbf{q}^1) < \Psi(\mathbf{q}^2)$ whenever two vectors \mathbf{q}^1 and \mathbf{q}^2 satisfy $\mathbf{q}^1 < \mathbf{q}^2$. (The inequalities are component-wise.) We now show that best-response pricing converges to the unique equilibrium. We define the sequence $\{\mathbf{q}^{(0)}, \mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \dots\} \subset \mathcal{Q}$ by $\mathbf{q}^{(k+1)} := \Psi(\mathbf{q}^{(k)})$ for $k \geq 0$.

Theorem 2. If each firm employs the best response strategy based on the prices of other firms in the previous iteration, the sequence of price vectors converges to the unique equilibrium price vector.

Proof. Let $\mathbf{q}^{(0)} \in \mathcal{Q}$ denote the attraction vector associated with the initial price vector. Choose $\underline{\mathbf{q}}^{(0)}, \overline{\mathbf{q}}^{(0)} \in \mathcal{Q}$ such that $\underline{\mathbf{q}}^{(0)} < \hat{b}^{(0)} < \overline{\mathbf{q}}^{(0)}$ and $\underline{\mathbf{q}}^{(0)} < \hat{b}^* < \overline{\mathbf{q}}^{(0)}$. Such $\underline{\mathbf{q}}^{(0)}$ and $\overline{\mathbf{q}}^{(0)}$ exist since \mathcal{Q} is a box-shaped open set.

For each $k \geq 0$, we define $\underline{\mathbf{q}}^{(k+1)} := \Psi(\underline{\mathbf{q}}^{(k)})$ and $\overline{\mathbf{q}}^{(k+1)} := \Psi(\overline{\mathbf{q}}^{(k)})$. From the monotonicity of $\Psi(\cdot)$ (Proposition 5) and $\Psi(\hat{b}^*) = \hat{b}^*$, we get

$$\underline{\mathbf{q}}^{(k)} < \hat{b}^{(k)} < \overline{\mathbf{q}}^{(k)} \quad \text{and} \quad \underline{\mathbf{q}}^{(k)} < \hat{b}^* < \overline{\mathbf{q}}^{(k)}. \tag{6}$$

Let $u^{(k)} := \max_i \left\{ \frac{\hat{q}_i^{(k)}}{\hat{q}_i^*} \right\}$. Clearly, $u^{(k)} > 1$ for all k by (6). We show that the sequence $\{u^{(k)}\}_{k=0}^\infty$ is strictly decreasing. For each i ,

$$\overline{q}_i^{(k+1)} = \psi_i(\hat{\overline{q}}_i^{(k)}) < \left(\hat{\overline{q}}_i^{(k)} + \kappa \right) \cdot \theta_i^* = \left(\hat{\overline{q}}_i^{(k)} + \kappa \right) \cdot \frac{q_i^*}{\hat{q}_i^* + \kappa} \leq \frac{\hat{\overline{q}}_i^{(k)}}{\hat{q}_i^*} \cdot q_i^* \leq u^{(k)} q_i^*,$$

where the first inequality comes from Proposition 6, the second one from (6), and the last one from the definition of $u^{(k)}$. Thus

$$\hat{\overline{q}}_i^{(k+1)} = \sum_{j \neq i} \overline{q}_j^{(k+1)} < \sum_{j \neq i} u^{(k)} q_j^* = u^{(k)} \hat{q}_i^*, \tag{7}$$

and $u^{(k+1)} = \max_i \{\hat{q}_i^{(k+1)} / \hat{q}_i^*\} < u^{(k)}$. Since $\{u^{(k)}\}_{k=0}^\infty$ is a monotone and bounded sequence, it converges. Let $u^\infty := \lim_{k \rightarrow \infty} u^{(k)}$. We claim $u^\infty = 1$. Suppose, by way of contradiction, $u^\infty > 1$. By Proposition 6, $\psi_i(\hat{q}_i) / (\hat{q}_i + \kappa)$ is strictly decreasing in \hat{q}_i . Thus, for any $\hat{q}_i \geq \frac{1}{2}(1 + u^\infty) \cdot \hat{q}_i^*$, there exists $\epsilon \in (0, 1)$ such that for each i , we have

$$\frac{\psi_i(\hat{q}_i)}{(\hat{q}_i + \kappa)} \leq (1 - \epsilon) \cdot \frac{\psi_i(\hat{q}_i^*)}{\hat{q}_i^* + \kappa} = (1 - \epsilon) \cdot \frac{q_i^*}{\hat{q}_i^* + \kappa} .$$

For any k , if $\hat{q}_i^{(k)} \geq \frac{1}{2}(1 + u^\infty) \cdot \hat{q}_i^*$, then

$$\begin{aligned} \hat{q}_i^{(k+1)} = \psi_i(\hat{q}_i^{(k)}) &\leq (\hat{q}_i^{(k)} + \kappa) \cdot (1 - \epsilon) \cdot \frac{q_i^*}{\hat{q}_i^* + \kappa} \\ &\leq \frac{\hat{q}_i^{(k)}}{\hat{q}_i^*} \cdot q_i^* \cdot (1 - \epsilon) \leq (1 - \epsilon) \cdot u^{(k)} \cdot q_i^* . \end{aligned}$$

Otherwise, we have $\hat{q}_i^* < \hat{q}_i^{(k)} < \frac{1}{2}(1 + u^\infty) \cdot \hat{q}_i^*$. By Proposition 6,

$$\hat{q}_i^{(k+1)} = \psi_i(\hat{q}_i^{(k)}) < (\hat{q}_i^{(k)} + \kappa) \cdot \frac{q_i^*}{\hat{q}_i^* + \kappa} \leq \frac{\hat{q}_i^{(k)}}{\hat{q}_i^*} \cdot q_i^* \leq \frac{1}{2}(1 + u^\infty) \cdot q_i^* .$$

Therefore, we conclude, using an argument similar to (7), $u^{(k+1)} \leq \max\{(1 - \epsilon) \cdot u^{(k)}, (1 + u^\infty) / 2\}$. From $u^{(k+1)} > u^\infty > (1 + u^\infty) / 2$, we obtain $u^{(k+1)} \leq (1 - \epsilon) \cdot u^{(k)}$, implying $u^{(k)} \rightarrow 1$ as $k \rightarrow \infty$. This is a contradiction. Similarly, we can show that $l^{(k)} := \min_i \{\hat{q}_i^{(k)} / \hat{q}_i^*\}$ is a strictly increasing sequence converging to 1. \square

The following proposition shows the linear convergence of tatônnement in the space of attraction values.

Proposition 7. *The sequence $\{\mathbf{q}^{(k)}\}_{k \geq 0}$ converges linearly.*

Proof. Consider $\{\underline{\mathbf{q}}^{(k)}\}_{k=0}^\infty$ and $\{\overline{\mathbf{q}}^{(k)}\}_{k=0}^\infty$ in the proof of Theorem 2. Recall $\underline{\mathbf{q}}^{(k)} < \hat{b}^{(k)} < \overline{\mathbf{q}}^{(k)}$ and $\underline{\mathbf{q}} < \hat{b}^* < \overline{\mathbf{q}}$. We will show that $\underline{\mathbf{q}}^{(k)}$ and $\overline{\mathbf{q}}^{(k)}$ converges to \hat{b}^* linearly. Since \mathcal{Q} is a box-shaped open set, there exists a convex compact set $\mathcal{B} \subset \mathcal{Q}$ containing all elements of $\{\underline{\mathbf{q}}^{(k)}\}_{k=0}^\infty$ and $\{\overline{\mathbf{q}}^{(k)}\}_{k=0}^\infty$. From the proof of Proposition 6, there exists $\delta > 0$ such that for any $\hat{b} \in \mathcal{B}$, we have

$$\frac{d}{d\hat{q}_i} \left(\frac{\psi(\hat{q}_i)}{\hat{q}_i + \kappa} \right) \leq -\delta .$$

From integrating both sides of the above expression from \hat{q}_i^* to $\hat{q}_i^{(k)}$,

$$\frac{\psi_i(\hat{q}_i^{(k)})}{\hat{q}_i^{(k)} + \kappa} - \frac{\psi_i(\hat{q}_i^*)}{\hat{q}_i^* + \kappa} \leq -\delta (\hat{q}_i^{(k)} - \hat{q}_i^*)$$

since the line segment connecting \hat{b}^* and $\overline{\mathbf{q}}^{(k)}$ lies within \mathcal{B} .

Define $\delta_1 := \delta \cdot \min_i \min_{\hat{q}_i \in \mathcal{B}} \{(\hat{q}_i + \kappa) \cdot \hat{q}_i^*/q_i^*\} > 0$. We choose $\delta > 0$ is sufficiently small such that $\delta_1 < 1$. Recall $\bar{q}_i^{(k+1)} = \psi_i(\hat{q}_i^{(k)})$ and $q_i^* = \psi(\hat{q}_i^*)$. Rearranging the above inequality and multiplying it by $(\hat{q}_i^{(k)} + \kappa)/q_i^*$,

$$\begin{aligned} \bar{q}_i^{(k+1)}/q_i^* &\leq (\hat{q}_i^{(k)} + \kappa)/(\hat{q}_i^* + \kappa) - \delta \cdot (\hat{q}_i^{(k)} - \hat{q}_i^*) \cdot (\hat{q}_i^{(k)} + \kappa)/q_i^* \\ &\leq \hat{q}_i^{(k)}/\hat{q}_i^* - \delta_1 \cdot (\hat{q}_i^{(k)}/\hat{q}_i^* - 1) = (1 - \delta_1) \cdot \hat{q}_i^{(k)}/\hat{q}_i^* + \delta_1. \end{aligned}$$

where the second inequality comes from $\hat{q}_i^{(k)} > \hat{q}_i^*$ and the definition of δ_1 .

Let $\rho(k) := \max_i \{\bar{q}_i^{(k)}/q_i^*\}$. Thus, $\bar{q}_j^{(k)} \leq \rho(k) \cdot q_j^*$ holds for all j , and summing this inequality for all $j \neq i$, we get $\hat{q}_i^{(k)} \leq \rho(k) \cdot \hat{q}_i^*$. Thus, $\bar{q}_i^{(k+1)}/q_i^*$ is bounded above by $(1 - \delta_1) \cdot \rho(k) + \delta_1$ for each i , and we obtain $\rho(k+1) \leq (1 - \delta_1) \cdot \rho(k) + \delta_1$. Using induction, it is easy to show

$$\rho(k) \leq (1 - \delta_1)^k \cdot (\rho(0) - 1) + 1$$

Therefore, we obtain

$$\begin{aligned} \max_i \left\{ (\bar{q}_i^{(k)} - q_i^*)/q_i^* \right\} &= \rho(k) - 1 \leq (1 - \delta_1)^k \cdot (\rho(0) - 1) \\ &= (1 - \delta_1)^k \cdot \max_i \left\{ (\bar{q}_i^{(0)} - q_i^*)/q_i^* \right\} \quad \text{and} \\ \max_i \left\{ \bar{q}_i^{(k)} - q_i^* \right\} &\leq (1 - \delta_1)^k \cdot \max_i \{q_i^*\} \cdot \max_i \left\{ (\bar{q}_i^{(0)} - q_i^*)/q_i^* \right\}, \end{aligned}$$

showing the linear convergence of the upper bound sequence $\bar{\mathbf{q}}^{(k)}$. A similar argument shows the linear convergence of $\underline{\mathbf{q}}^{(k)}$. □

The linear convergence in the above theorem is not with respect to the price vector but with respect to the attraction vector, i.e. not in \mathbf{p} but in \mathbf{q} . Yet, the following theorem also shows the linear convergence with respect to the price vector. Let $\{\mathbf{p}^{(k)}\}_{k \geq 0}$ be the sequence defined by $p_i^{(k)} := a_i^{-1}(q_i^{(k)})$.

Theorem 3. *The sequence $\{\mathbf{p}^{(k)}\}_{k \geq 0}$ converges linearly.*

Proof. Consider $\{\underline{\mathbf{q}}^{(k)}\}_{k=0}^\infty$ and $\{\bar{\mathbf{q}}^{(k)}\}_{k=0}^\infty$ in the proof of Proposition 7. Let $\{\underline{\mathbf{p}}^{(k)}\}_{k=0}^\infty$ and $\{\bar{\mathbf{p}}^{(k)}\}_{k=0}^\infty$ be the corresponding sequences of price vectors.

Within the compact interval $[\bar{p}_i^{(0)}, p_i^{(0)}]$, the derivative of a_i is continuous and its infimum is strictly negative. By the Inverse Function Theorem, the derivative of $a_i^{-1}(\cdot)$ is continuous in the compact domain of $[q_i^{(0)}, \bar{q}_i^{(0)}]$. Recall $p_i^* = a_i^{-1}(q_i^*)$. There exists some bound $M > 0$ such that

$$|p_i - p_i^*| = |a_i^{-1}(q_i) - a_i^{-1}(q_i^*)| \leq M \cdot |q_i - q_i^*|$$

whenever $a_i(p_i) = q_i \in [q_i^{(0)}, \bar{q}_i^{(0)}]$ for all i . From the proof of Proposition 7, we obtain $q_i^{(k)} \in (q_i^{(k)}, \bar{q}_i^{(k)}) \subset [q_i^{(0)}, \bar{q}_i^{(0)}]$. Therefore, the linear convergence of $\{\mathbf{q}^{(k)}\}_{k \geq 0}$ implies the linear convergence of $\{\mathbf{p}^{(k)}\}_{k \geq 0}$. □

References

1. G. Allon and A. Federgruen. Service competition with general queuing facilities. Technical report, Working Paper, 2004.
2. S. P. Anderson, A. De Palma, and J.-F. Thisse. *Discrete Choice Theory of Product Differentiation*. The MIT Press, 1996.
3. F. Bernstein and A. Federgruen. Comparative statics, strategic complements and substitutes in oligopolies. *Journal of Mathematical Economics*, 40(6):713–746, 2004.
4. F. Bernstein and A. Federgruen. A general equilibrium model for industries with price and service competition. *Operations Research*, 52(6):868–886, 2004.
5. D. Besanko, S. Gupta, and D. Jain. Logit demand estimation under competitive pricing behavior: An equilibrium framework. *Management Science*, 44:1533–1547, 1998.
6. G. P. Cachon and P. T. Harker. Competition and outsourcing with scale economies. *Management Science*, 48(10):1314–1333, 2002.
7. A. Caplin and B. Nalebuff. Aggregation and imperfect competition: On the existence of equilibrium. *Econometrica*, 59:25–59, 1991.
8. E. H. Chamberlin. *The Theory of Monopolistic Competition*. Harvard University Press, 1933.
9. P. Dubey, O. Haimanko, and A. Zapechelnuk. Strategic complements and substitutes, and potential games. Technical report, Working Paper, 2003.
10. J. Friedman. *Oligopoly Theory*. Cambridge University Press, 1983.
11. G. Gallego, W. T. Huh, W. Kang, and R. Philips. Price Competition with Product Substitution: Existence of Unique Equilibrium and Its Stability. Technical report, Working Paper, 2005.
12. H. Hotelling. Stability in competition. *Economic Journal*, 39:41–57, 1929.
13. R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.
14. S. Mahajan and G. Van Ryzin. Supply chain contracting and coordination with stochastic demand. In S. Tayur, M. Magazine, and R. Ganeshan, editors, *Quantitative Models for Supply Chain Management*. Kluwer, 1998.
15. J. Medhi. *Stochastic Models in Queuing Theory*. Academic Press, second edition, 2003.
16. P. Milgrom and J. Roberts. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica*, 58:1255–1277, 1990.
17. P. Milgrom and C. Shannon. Monotone comparative statics. *Econometrica*, 62:157–180, 1994.
18. T. Mizuno. On the existence of a unique price equilibrium for models of product differentiation. *International Journal of Industrial Organization*, 62:761–793, 2003.
19. J. Roberts and H. Sonnenschein. On the foundations of the theory of monopolistic competition. *Econometrica*, 45:101–113, 1977.
20. J. Robinson. *The Economics of Imperfect Competition*. Macmillan, 1933.
21. K. C. So. Price and time competition for service delivery. *Manufacturing & Service Operations Management*, 2:392–407, 2000.
22. D. M. Topkis. Equilibrium points in nonzero-sum n -person submodular games. *SIAM Journal on Control and Optimization*, 17:773–787, 1979.
23. X. Vives. Nash equilibrium with strategic complementarities. *Journal of Mathematical Economics*, pages 305–321, 1990.
24. X. Vives. *Oligopoly Pricing: Old Ideas and New Tools*. The MIT Press, 1999.

Design for Using Purpose of Assembly-Group

Hak-Soo Mok¹, Chang-Hyo Han¹, Chan-Hyoung Lim², John-Hee Hong³,
and • Jong-Rae Cho³

¹ Dept. of Industrial Engineering, Pusan National University,
30 Changjeon-Dong, Kumjeong-Ku, Busan, 609-735, Korea
Tel.: (051)510-1435; Fax: (051)512-7603
hsmok@pusan.ac.kr

² Nemo Solutions Consulting, 7th floor, Korea Electric Power Corporation Building, 21,
Yeoido-dong, Yeongdeungpo-gu, Seoul, 150-875, Korea
Tel.: (02)3454-0340; Fax: (02)786-0349
chan0070@hotmail.com

³ Hyundai Kia Motors co, Corporate Research & Development Division 104, Mabuk-Ri,
Guseong-Eup, Yongin-Si, Gyeonggi-Do, 449-912, Korea
Tel.: (031) 899-3056; Fax: (031)368-6786
muyoung@hyundai-motor.com

Abstract. In this paper, the disassemblability is determined by the detail weighting factors according to the using purpose of assembly-group. Based on the disassembly mechanism and the characteristics of parts and assembly-groups, the disassembly functions are classified into three categories; accessibility, transmission of disassembly power and disassembly structure. To determine the influencing parameters, some assembly-groups of an automobile are disassembled. The weighting values for the influencing factors are calculated by using of AHP (Analytic Hierarchy Process). Using these weighting values, the point tables for the using purpose are determined. Finally, an optimal design guideline for the using purpose of an assembly-group can be decided.

1 Introduction

The shortage of landfill and waste burning facilities constantly reminds us that our products do not simply disappear after disposal. It is currently acknowledged that the most ecologically sound way to treat a worn out product is recycling. Because disassembly is related to recycling and is a necessary and critical process for the end-of life (EOL) of a product, the design methodology should be developed in terms of environmentally conscious designs (Eco-design). Disassembly can be defined as a process of systematic removal of desirable parts from an assembly while ensuring that parts are not impaired during the process.[1] The goal of disassembly for recycling is to separate different materials with less effort. There should be a countermeasure for companies to reduce the recycling expenses. There is also increased demand for products that can be easily maintained. In other words, by the reducing the disassembly time we can decrease the man-hours. Now the environmental problems are seriously discussed among many companies. Fig. 1 shows the life cycle of a worn

out product. In order to design a product, which is environmentally benign, the life cycle of the product should be well understood. The disassembly technology should be systematized to reduce the recycling cost and time, because the worn out products transported from the logistics center can be reused or recycled after disassembly processes.[2],[3].

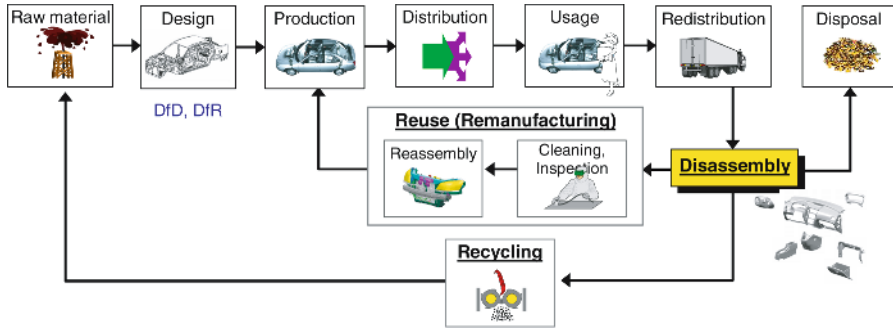


Fig. 1. Location of disassembly in life cycle of a product

The EU policy is calling for mandatory environmental policies for EU-member countries to regulate scrapped vehicles recycling. From early 2006, 85% of the weight of a scrapped vehicle should be recycled and 95% including 10% energy of a vehicle should be recycled after 9 years. [4],[5]

In order to recycle an assembled good, there must be disassembly and sorting processes. To conduct these processes easily, the structure of a product and assembly-groups should be oriented to disassembly. Furthermore, the product and assembly-groups have to be designed with consideration of their using purposes.

In this paper, the using purposes are divided into (1) user aspect, (2) A/S aspect, (3) reuse aspect and (4) recycling aspect. According to these four categories, a new product and assembly-groups should be designed. Then, the main activities for this research are showed in the followings:

- * Analysis of the mechanism of disassembly and understanding of the weak disassembly processes
- * Determination of the influencing parameters on the disassembly
- * Determination of the detail weighting factors for detail disassembly functions
- * Determination of the weighting values for using purposes of assembly-groups
- * Evaluation of the disassemblability points
- * Establishment of the point tables of the disassemblability
- * Selection of an optimal alternative among several design guidelines

2 Detail Functions of Disassembly Definition of Disassemblability

In this paper, we determined that five detail functions are normally needed for disassembly: fixing of the object, accessing the disassembly point, transmitting the

disassembly power, grasping and handling of the object. Fig. 2 shows the definitions of these detail functions. In order to improve the disassemblability of a product, the previous five detail functions should be simplified and be conducted more easily.

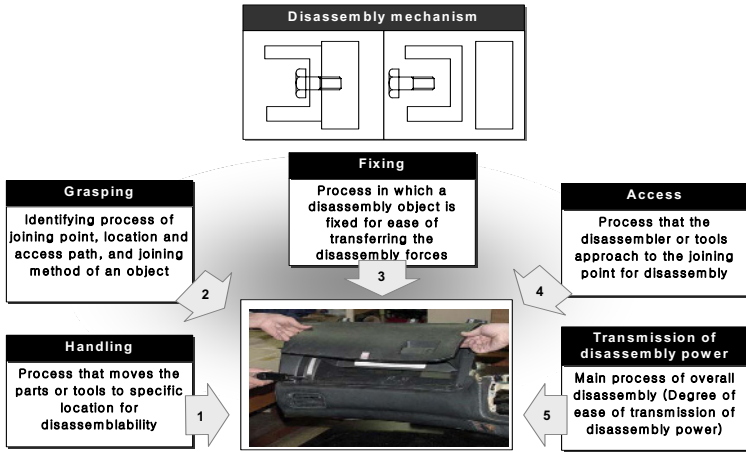


Fig. 2. Definition of detail functions of disassembly

In Fig. 3, the detail functions of disassembly are classified according to the required disassembly time and the object of the assembly-group [6].

Before modification		Disassembly of subassemblies	Disassembly of parts
	Main-process	<ul style="list-style-type: none"> • Grasping • Access process • Transmission of disassembly power 	<ul style="list-style-type: none"> • Fixing • Access • Handling • Transmission of disassembly power
After modification		Disassembly of subassemblies	Disassembly of parts
	Main-process	<ul style="list-style-type: none"> • Access (including grasping) • Transmission of disassembly power 	<ul style="list-style-type: none"> • Access (including handling, fixing) • Transmission of disassembly power (including fixing)
	Structure feature	<ul style="list-style-type: none"> • Part structure (part number of pre-disassembled, arrangement of the parts) • Assembly structure (number of assembly factors, number of connecting parts, number of joining points) 	

Fig. 3. Classification of the disassembly functions

The object of disassembly-groups is divided into the disassembly of product and the disassembly of part. In the disassembly of product, it is not needed the fixing processes, because the product is fixed and stable by the weight of the product. In this case, only three detail functions would be used: (1) access into the object, (2) transmission of the disassembly power and (3) structure of the product (Table. 1).[7]

Table 1. Definition of detail disassembly function

	Category	Definition
Process	accessibility	Degree of ease of access to joining point or specification position for disassembly.
	Transmission of disassembly power	Degree of ease of transmit the disassembly power at disassembly position
Structure	Disassembly Structure	Structural properties of objects that influence on diassemblability

For the analysis of these disassembly functions, the geometrical properties of parts, the applied fastening methods, the connecting parts and the number of parts are checked.

3 Determination of the Detail Weighting Factors

In order to determine the weighting factors of the detail disassembly function for the product and the assembly-group, several disassembly experiments were performed in the lab. Using these experimental data, the criteria for the evaluation of the diassemblability and the levels for each criterion should be determined. The difficulty of disassembly function, the needed personal motion, the average disassembly time, the frequency of disassembly process and the frequency of weak process are considered as the criteria.

The evaluation level is determined by the analysis of the disassembly experiment of assembly-groups. Using these criteria and the levels, we can find the weighting factors for the detail disassembly process. Table 2 shows the procedure to find the detail weighting factors. The detail weighting factor for the accessing process is the value in 0.342 and for the transmission of disassembly power is 0.658. If the weighting factor for the structural property will be 0.5, the weighting factor for accessibility will be 0.171 and the weighting factor for the transmission of disassembly power is 0.329. Because the characteristics of product structure play an important role in the disassembly process.

Table 2. Detail weighting factors of detail diassemblability

	Evaluation criteria					Total	
	Difficulty of process	Grade of necessary in Body part	Average work time	Frequency of process	Frequency of bottleneck process		
Accessibility	3	3	1	3	3	13/38 =0.342	<ul style="list-style-type: none"> • The weight of accessibility: $0.342 * 0.5 = 0.171$ • The weight of transmission of disassembly power : $0.658 * 0.5 = 0.329$ • The weight of disassembly structure : 0.500
Transmission of disassembly power	5	5	5	5	5	25/38 =0.658	

4 Determination of an Optimal Alternative of Design Guideline Using the Point Table for Disassembly

The following steps are used to determine the point of disassemblability:

- Determination of the detail weighting factors ; (in 3chapter)
- Determination of the weighting values(w_i) for the influencing parameters
- Multiplication of the detail weighting factor(w_j) for the disassemblability by the weighting value for the influencing parameter ; (This equals total weighting factor)
- Multiplication of the score for the level of the detail influencing parameter for the detail influencing parameter by the total weighting factor

The disassembly experiments are performed to determine the detail weighting factors, which are used to identify the weak disassembly process and more complex parts that cannot be disassembled easily. Then using these results, the detail weighting factors for the detail influencing parameters can be calculated by the AHP (Analytic Hierarchy Process).

The total weighting factor is a main contribution of the higher quantitative evaluation of the disassemblability. The score for the level of the detail influencing parameter can be decided by the given disassembly conditions and the level number.

The following step is the procedure to estimate the disassemblability points for the disassembly purpose according to the user of a product.

<p>Step 1. Definition the weighting factors(w_i) for the detail disassemblability (ex. the weighting for accessibility : 0.171)</p> <p>Step 2. Definition the weighting value(w_j) for each influencing parameter</p> <ul style="list-style-type: none"> • Size of access space : 0.281 • Visuality of access route : 0.260 • Self-location : 0.299 • Number of access direction : 0.160 <p>Step 3. Multiplication $\langle w_i * w_j * 100 \rangle$ (The weight of access space: $0.171 * 0.281 * 100$)</p> <p>Step 4. Multiplication the detail disassemblability for each influencing parameter and the score value of Step 3. (here, the detail disassemblability with 3 grade : 1 , 3, 5 point and 2 grade : 1, 5 point)</p> <p>Step 5. As the disassemblability point is determined the sum of Step 4 for each influencing parameter</p>

Using the point of the disassemblability, the point table of the disassemblability for each using purpose of the product can be established. In the point table of the disassemblability we can find the position according to the given disassembly condition: the property of access, the visual ability, the existence of self-location and the number of access directions. From this position we can have four possible alternatives to improve the disassemblability of a product. Here we assume that it is possible that an alternative can be obtained by the changing only one parameter at a time. This assumption could give us the simple solution.

This suggests that the possible direction of the improvement can not be found in a diagonal line. Now in this step, an optimal alternative can be determined by choosing the alternative that has the highest score of the disassemblability.

Fig. 4 shows the step used to determine the possible alternative of the design guideline. In the first step, the disassemblability point is determined according to the disassembly condition. Then, we can find the position in the point table. For example, the disassemblability point is 41 in the given disassembly condition: when the access is direction-restricted, a visibility is medium, the selflocation is zero, and the change of disassembly direction is one. In the second step, the alternative is determined using our algorithm to get a design guideline. In this paper, there are four alternatives.

In the third step, an optimal design alternative is determined from the comparison of the point difference among the occurred alternatives. In this case, the fourth alternative is selected as the best solution from among four alternatives, because the rising gap of the score of the disassemblability is the highest point (here, 21 point). In other words, the most valuable design guideline is: a self-location should be established in the product.

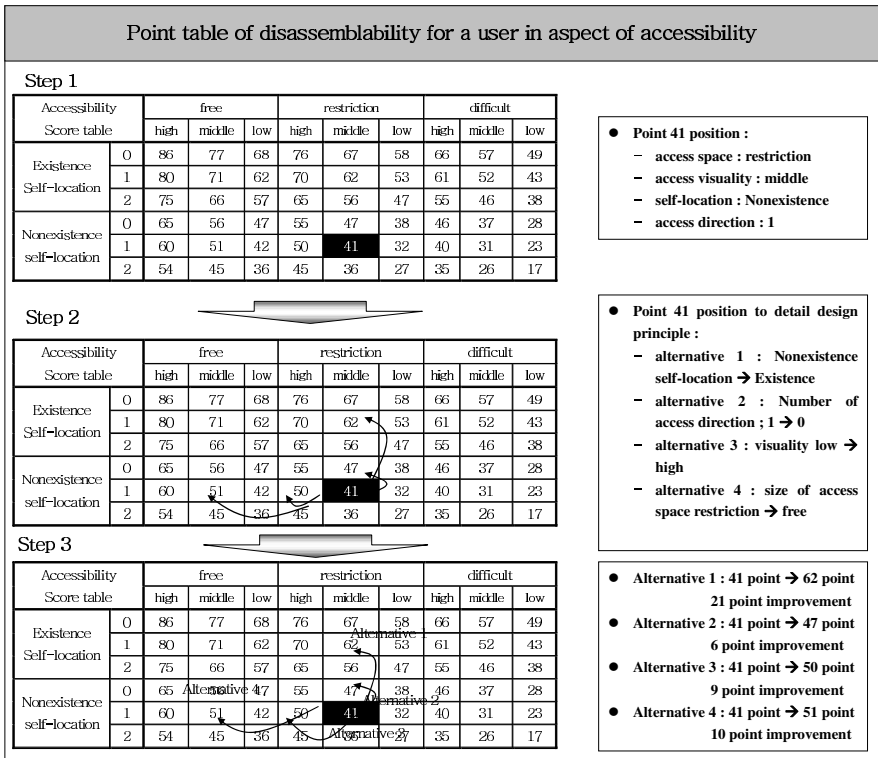


Fig. 4. The point table of disassemblability for the user according to the accessibility

5 Case Study

As a case study we considered the air cleaner of an automobile. By disassembling air cleaner, we found several weak disassembly processes. The condition of the disassembly process and the characteristics of three assembly-groups are analyzed according to their using purpose. Fig. 5 shows the disassembly condition and the position in the point table. And we can have four possible design guidelines in the aspects of a user of product.

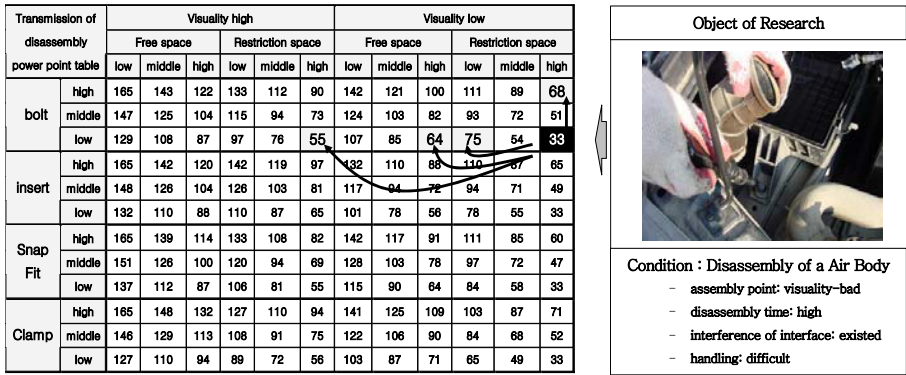


Fig. 5. The current status of the transmission of disassembly power in the aspects of User and the design principle

Table 3. The design principle in the purpose of usage from disassemblability

Design of principle	Detail disassemblability		Accessibility	Transmission of disassembly power	Disassembly structure
	Purpose of using				
Design of principle	User	● Existence self-location	● Valuality high	● Connect part add	
	A/S	● Existence self-location	● Disassembly power down	● Connect part add	
	Reuse	● Existence self-location	● Grasp-ability high	● Assembly point down	
	Recycling	● Existence self-location	● Disassembly power down	● Assembly point down	

In this case study, the disassembly conditions are: the assembly method is bolting, the gripping ability is low, the visual ability is low, and the working area is not enough (limited), the disassembly force is high needed. Based on these experimented data, we could determine the position in the point table of the disassemblability. The disassemblability point is 33 for the user-oriented purpose.

In this position, we could find four possible design guidelines. From among these four alternatives we choose one with the largest value (here, 75). It is the optimal design guideline in this case.

In order to prepare an alternative for a design guideline of subassembly (air cleaner), some design guidelines according to the purpose of usage are shown in Table 3. In this case, its visibility is low at the point of bolting to the body, transmission of disassembly power is in low condition, also the visibility of the parts of duct is low and assembly point is hard to confirm. Since it was difficult to access the disassembly point with tools, the transmission of disassembly power was low. Fig. 6 shows an alternative of the design guideline from the aspects of the user.



User		Principle: Reduce the disassembly power	
Disassembly process : Air Body		Alternative	
			
Assembly factor	Disassembly time	Assembly factor	Disassembly time
Re-bolt	19.68 sec	Re-bolt	12.27 sec
Problem			
<ul style="list-style-type: none"> • Assembly point: Not visible • Disassembly time: Much • Interference: Exist • Handling: Bad 			
Alternative			
<ul style="list-style-type: none"> • Remove the interference • Move the assembly point • Reduce the number of assembly factors 			

Fig. 6. Design guidelines in the aspects of User for Duct and Air Body

6 Conclusion

In this paper, the using purpose of an assembly-group is divided into four categories: aspect of user, A/S, reuse and recycling. According to these categories, the detail disassembly function and the influencing parameters were considered. The total weighting factors were used to evaluate the disassemblability.

To choose an optimal alternative for a disassembly-oriented design we established the point table of disassemblability. Using the point table we found the position. This position shows the quantitative disassemblability for a given disassembly condition. The suggested algorithm was used to find an optimal design guideline systematically.

Acknowledgment

This work was supported by Cleaner Production Technology Development of Ministry of Commerce, Industry and Energy (20040521) and Pusan National Research Grant.

References

1. Mok, H. S, Moon, K. S, Kim, S. H, Moon, D, S.: The Complexity Evaluation System of Automobile Subassembly for Recycling, Journal of the Korean society of Precision Engineering 16(5), (1999) 132-144
2. Mok, H. S, Kim, S. H, Yang, T. I.: Evaluation of the Product Complexity Considering the Disassemblability, Journal of the Korean society of Precision Engineering 16(5), (1999) 14-24
3. Mok, H. S, Cho, J. R.: Development of Product Design Methodology for Assemblability and Disassemblability Considering Recycling, Journal of the Korean society of Precision Engineering 18(7), (2001) 72-84
4. Autorecycling Wohin fuehrt der Weg, RECYCLING magazin, 4, (2003)
5. European Union, Speeding up the vehicle recycling, EU Kommission, 3 (2003)
6. Moon, K. S.: Analysis and Evaluation System for Disassembly, Ph.D thesis, Pusan... National University, Busan, South Korea, (2003)
7. Mok H. S, Hwang Hoon, Yang Tae-II.: Module design of a product considering disassemblability, Korean Institute of Industrial Engineers/The Korean Operations and Management Science Society collaborated art and science contest in, Spring. A collection of a treaties. (2000)

A Conditional Gaussian Martingale Algorithm for Global Optimization

Manuel L. Esquível*

Departamento de Matemática and CMA, FCT/UNL, Quinta da Torre,
2829-516 Caparica, Portugal

mle@fct.unl.pt

<http://ferrari.dmat.fct.unl.pt/personal/mle/>

Abstract. A new stochastic algorithm for determination of a global minimum of a real valued continuous function defined on K , a compact set of \mathbb{R}^n , having an unique global minimizer in K is introduced and studied, a context discussion is presented and implementations are used to compare the performance of the algorithm with other algorithms. The algorithm may be thought to belong to the *random search* class but although we use Gaussian distributions, the mean is changed at each step to be the intermediate minimum found at the preceding step and the standard deviations, on the diagonal of the covariance matrix, are halved from one step to the next. The convergence proof is simple relying on the fact that the sequence of intermediate random minima is an uniformly integrable conditional Gaussian martingale.

1 Introduction

Quite some attention has been recently devoted to stochastic algorithms, as more than 300 bibliographic entries in the reference textbook [9] testifies. Highly schematized global optimization methods using randomized search strategies are object of a thorough synthetic theoretical study in [12] which also presents applications of these methods to engineering problems. Negative results as those in [10] show that overconfidence on the effectiveness of stochastic methods is not desirable but, nevertheless, it is natural to speculate that an adequate randomized algorithm may perform better than a deterministic one in global optimization, at least in most of the situations. Theoretical results such as those in [2], [11] and [7], indicate that stochastic algorithms may be thought to be as reliable as deterministic ones and efforts in order to find better performing algorithms continue to be pursued as in [1] and [6]. The main feature of the new algorithm presented here, allows to recover some interesting properties of other stochastic algorithms such as the clustering and adaptiveness properties simultaneously with the property of continuing to search the whole domain at each step, which is a characteristic feature of simulated annealing.

* This work was done with partial support of FCT (Fundação para a Ciência e Tecnologia) program POCTI (Portugal/FEDER-EU). I hereby express my gratitude to Ana Luísa Custódio for enlightening discussions and suggestions and to the Mathematics Department for allowing an intensive use of one of its computer laboratories.

2 The Solis and Wets Approach to Random Search

We recall next the powerful meta-approach of Solis and Wets (see [8]) in order to generalize its formulation to the case of adaptive random search and almost sure convergence. The original formulation of the authors solves the problem for non adaptive random search and convergence in probability, as noted in the remark 1 ahead. According to [2, p. 22], with the exception of [8] there were no synthesis studies of random search algorithms prior to 1986. Consider $f : K \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$, K a Borel set, where we suppose that for $x \in K^c$ we have $f(x) = +\infty$ and, $(\Omega, \mathcal{A}, \mathbb{P})$ a complete probability space. The following general selection scheme is the nuclear part of the random algorithm. Let $\psi : K \times \mathbb{R}^n \rightarrow K$ be such that the following hypothesis [H1] is verified.

$$\begin{cases} \forall x, t & f(\psi(t, x)) \leq f(t) \\ \forall x \in K & f(\psi(t, x)) \leq f(x) . \end{cases} \tag{2.1}$$

The general conceptual algorithm of Solis and Wets is as follows.

S. 0: Take $t_0 \in K$ and set $j = 0$.

S. 1: Generate a point x_j from $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}_j)$.

S. 2: Set $t_{j+1} = \psi(t_j, x_j)$ choose \mathbb{P}_{j+1} , set $j = j + 1$ and return to step 1 (S.1).

Observe that in adaptive methods, x_j is a point with distribution \mathbb{P}_j which depends on $t_{j-1}, t_{j-2}, \dots, t_0$, thus being a conditional distribution. Let now λ denote the Lebesgue measure over $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and α be the essential infimum of f over K , that is: $\alpha := \inf\{t \in \mathbb{R} : \lambda(\{x \in K : f(x) < t\}) > 0\}$. Consideration of the essential infimum is mandatory to deal correctly with non continuous or unbounded functions such as $\mathbb{1}_{[0,1] \setminus \{1/2\}}$ defined in $[0, 1]$, or $\ln(|x|)\mathbb{1}_{\mathbb{R} \setminus \{0\}} + (-\infty)\mathbb{1}_{\{0\}}$. Let $E_{\alpha+\epsilon, M}$ denote the level set of f having level $\alpha + \epsilon$ defined by:

$$E_{\alpha+\epsilon, M} := \begin{cases} \{x \in K : f(x) < \alpha + \epsilon\} & \text{if } \alpha \in \mathbb{R} \\ \{x \in K : f(x) < M\} & \text{if } \alpha = -\infty . \end{cases} \tag{2.2}$$

A Solis and Wets's type convergence theorem may now be formulated and proved.

Theorem 1. *Suppose that f is bounded from below. Let the sequence of random variables $(T_j)_{j \geq 0}$ be defined inductively by using the sequence $(X_j)_{j \geq 0}$ which, in turn, depends on the family of probability laws $(\mathbb{P}_j)_{j \geq 0}$ given by the algorithm above and verifying: $T_0 = X_0$ such that $X_0 \sim \mathbb{P}_0$, $X_j \sim \mathbb{P}_j$ (to mean that X_j has \mathbb{P}_j as law) and $T_{j+1} = \psi(T_j, X_j)$. If we have that the following hypothesis [H2(ϵ)] is verified: for some $\epsilon \geq 0$ and $M \in \mathbb{R}$*

$$\lim_{k \rightarrow +\infty} \inf_{0 \leq j \leq k-1} \mathbb{P}_j[E_{\alpha+\epsilon, M}^c] = \lim_{k \rightarrow +\infty} \inf_{0 \leq j \leq k-1} \mathbb{P}[X_j \in E_{\alpha+\epsilon, M}^c] = 0 , \tag{2.3}$$

then,

$$\lim_{k \rightarrow +\infty} \mathbb{P}[T_j \in E_{\alpha+\epsilon, M}] = 1 . \tag{2.4}$$

If, for all $\epsilon > 0$ [H2(ϵ)] is true, $(f(T_j))_{j \geq 0}$ converges almost surely to a random variable Y_* such that:

$$\mathbb{P}[Y_* \leq \alpha] = 1 . \tag{2.5}$$

Proof. Observe first that by hypothesis [H1] in formula 2.1 we have that if $T_k \in E_{\alpha+\epsilon, M}$ or $X_k \in E_{\alpha+\epsilon, M}$ then for all $n \geq 0$, $T_{k+n} \in E_{\alpha+\epsilon, M}$. As a consequence,

$$\{T_k \in E_{\alpha+\epsilon, M}^c\} \subseteq \{T_1, T_2, \dots, T_{k-1} \in E_{\alpha+\epsilon, M}^c\} \cap \{X_1, X_2, \dots, X_{k-1} \in E_{\alpha+\epsilon, M}^c\} .$$

So, for all $j \in \{0, 1, \dots, k - 1\}$:

$$\begin{aligned} \mathbb{P}[T_k \in E_{\alpha+\epsilon, M}^c] &\leq \mathbb{P} \left[\bigcap_{0 \leq l \leq k-1} \{T_l \in E_{\alpha+\epsilon, M}^c\} \cap \{X_l \in E_{\alpha+\epsilon, M}^c\} \right] \leq \\ &\leq \mathbb{P}[X_j \in E_{\alpha+\epsilon, M}^c] = \mathbb{P}_j[E_{\alpha+\epsilon, M}^c] , \end{aligned}$$

which implies $\mathbb{P}[T_k \in E_{\alpha+\epsilon, M}^c] \leq \inf_{0 \leq j \leq k-1} \mathbb{P}_j[E_{\alpha+\epsilon, M}^c]$. We may now conclude that:

$$1 \geq \mathbb{P}[T_k \in E_{\alpha+\epsilon, M}] = 1 - \mathbb{P}[T_k \in E_{\alpha+\epsilon, M}^c] \geq 1 - \inf_{0 \leq j \leq k-1} \mathbb{P}_j[E_{\alpha+\epsilon, M}^c] ,$$

which, as a consequence of formula 2.3 implies the conclusion in formula 2.4. Define now the filtration $\mathbb{G} = (\mathcal{G}_j)_{j \geq 0}$ by $\mathcal{G}_j = \sigma(T_0, T_1, \dots, T_j)$. It is then clear, from hypothesis [H1] in formula 2.1, that:

$$\mathbb{E}[f(T_{j+1}) \mid \mathcal{G}_j] = \mathbb{E}[f(\psi(T_j, X_j) \mid \mathcal{G}_j] \leq \mathbb{E}[f(T_j) \mid \mathcal{G}_j] = f(T_j) ,$$

thus showing that $(f(T_j)_{j \geq 0})$ is a supermartingale bounded from below which we know to be almost surely convergent to some random variable which we will denote by Y_* . This conclusion together with formula 2.4 already proved shows that formula 2.5 yields, as a consequence of lemma 1.

Remark 1. Hypothesis [H2] given by formula 2.3 mean that the more the algorithm progresses in its steps, the more mass of the distributions \mathbb{P}_j should be concentrated in the set $E_{\alpha+\epsilon, M}$ where the interesting points are. Our formulation of hypothesis [H2] differs from the one presented in [8] which reads:

$$\forall A \in \mathcal{B}(\mathbb{R}^n) \quad \lambda(A) = 0 \Rightarrow \prod_{j=0}^{+\infty} (1 - \mathbb{P}_j[A]) = 0 . \tag{2.6}$$

Formula 2.3 implies that, for $\epsilon > 0$, we have $\prod_{j=0}^{+\infty} (1 - \mathbb{P}_j[E_{\alpha+\epsilon, M}]) = 0$ and so, our hypothesis is stronger than the one in [8]. The hypothesis given by formula 2.6 is more appealing as it does not use a condition on the set $E_{\alpha+\epsilon, M}$ which, in general, is not explicitly known and, in almost every case, will be difficult to use computationally. On the other hand, hypothesis given by formula 2.6 does not allow the conclusion of the Convergence Theorem (Global Search) in [8, p. 20] to hold in full generality. The theorem is true, with the proof presented there, if the sequence $(X_j)_{j \geq 0}$ is a sequence of independent random variables. The authors do not mention this caveat and the phrase ... Nearly all random search methods

are adaptive by which we mean that μ_k^1 depends on the quantities ... generated by the preceding iterations ... *may induce the reader in the opposite belief. In fact, the inequality on the right in the third line in the proof of the theorem (see [8, p. 21]) is, in the general case of dependent draws of the $(X_j)_{j \geq 0}$, the reversed one as a consequence of the elementary fact that if $A \subset B$ and $0 < \mathbb{P}[B] < 1$ then: $\mathbb{P}[A \cap B] = \mathbb{P}[A] = \mathbb{P}[B] \cdot \mathbb{P}[A]/\mathbb{P}[B] \geq \mathbb{P}[A] \cdot \mathbb{P}[B]$.*

If f is continuous over K , a compact subset of \mathbb{R}^n , then f attains its minimum and it is verified that $\alpha = \min_{x \in K} f(x)$. The conceptual algorithm above furnishes a way of determining this minimum. This is a simple consequence of the following result, for which the proof is easily seen.

Corolary 1. *Under the same hypothesis, if in addition for all $x \in K$ we have $f(x) \geq \alpha$, then: $\mathbb{P}[Y_* = \alpha] = 1$.*

For the reader’s commodity, we state and prove the lemma used above.

Lemma 1. *Let $(Z_j)_{j \geq 0}$ be a sequence of random variables such that almost surely we have $\lim_{j \rightarrow +\infty} Z_j = Z$ and for some $\delta > 0$ we have $\lim_{j \rightarrow +\infty} \mathbb{P}[Z_j < \delta] = 1$. Then, $\mathbb{P}[Z \leq \delta] = 1$.*

Proof. It is a consequence of a simple observation following Fatou’s lemma.

$$\begin{aligned} \mathbb{P}[Z > \delta] &= \mathbb{P}[(\liminf_{j \rightarrow +\infty} Z_j) > \delta] = \mathbb{P}[\liminf_{j \rightarrow +\infty} \{Z_j > \delta\}] \leq \liminf_{j \rightarrow +\infty} \mathbb{P}[\{Z_j > \delta\}] \leq \\ &\leq \limsup_{j \rightarrow +\infty} \mathbb{P}[\{Z_j \geq \delta\}] = \lim_{j \rightarrow +\infty} \mathbb{P}[\{Z_j \geq \delta\}] = 0 . \end{aligned}$$

3 The Conditional Gaussian Martingale (CGM) Algorithm

The algorithm presented here may be included, on a first approximation, in the class of random search methods as this class *consists of algorithms which generate a sequence of points in the feasible region following some prescribed probability distribution or sequence of probability distributions*, according to [3, p. 835]. The main idea of the method studied here is to change, at each new step, the location and dispersion parameters of the probability Gaussian distribution in order to concentrate the points, from which the new intermediate minimum will be selected, in the region where there is a greater chance of finding a global minimum not precluding, however, a new intermediate minimum to be found outside this region. We use a sequence of Gaussian distributions taking at each step the mean equal to the intermediate minimum found in the preceding step and the standard deviation diagonal elements of the covariance matrix equal to half the ones taken in the preceding step. We now briefly describe the algorithm and after we will present the almost sure convergence result. The goal is to find a global minimum of a real function f defined over a compact set $K \subset \mathbb{R}^n$ having

¹ In our notation, the \mathbb{P}_j .

diameter c . From now on, $\mathcal{U}(K)$ will denote the uniform distribution over K and $\mathcal{N}(m, \sigma)$ denotes the Gaussian distribution with mean m and covariance matrix with equal diagonal elements σ . With the presentation protocol of [5] the algorithm is as follows.

- S. 0 Set $j = 0$;
- S. 1 Generate $x_1^0, x_2^0, \dots, x_N^0$ from the uniform distribution over the domain K .
- S. 2 Choose $t_0 = x_{i_0}^0$ such that $f(x_{i_0}^0)$ is equal to $\min\{f(x_i^0) : 1 \leq i \leq N\}$. Increment j .
- S. 3 Generate $x_1^j, x_2^j, \dots, x_N^j$ from the normal distribution $\mathcal{N}(t_{j-1}, c/2^j)$ having mean t_{j-1} and diagonal covariance matrix elements $c/2^j$, c being the diameter of K .
- S. 4 Choose $t_j = t_{i-1}$ if $f(t_{i-1})$ is strictly inferior to $\min\{f(x_i^j) : 1 \leq i \leq N\}$ and choose $t_j = x_{i_0}^j$ if $f(x_{i_0}^j)$ is less or equal to $\min\{f(x_i^j) : 1 \leq i \leq N\} \leq f(t_{i-1})$.
- S. 5 Perform a stopping test and then: either stop, or increment j and return to step 3 (S. 3).

Observe that steps 1 and 2 are useful in order to choose a starting point for the algorithm in K . The repetition of steps 3, 4 and 5, provide a sort of clustering of the random test points around the intermediate minimizer found at the preceding step while continuing to search the whole space.

This algorithm's core is easily implemented in a programming language allowing symbolic computation (all implementations presented in this text are

```

Jota = 400; Uba = {}; Ruba = {};
For[j = 1, j <= Jota, j++,
  Tuba = {0}; Luba = {};
  minimo = Module[
    {ptminX = xm, ptminY = ym, ptmaxX = xM, ptmaxY = yM,
     cont1 = 850, alea1, alea, T1, M1, Tes, eMes, NunOr},
    alea1 = Table[{Random[Real, {ptminX, ptmaxX}],
      Random[Real, {ptminY, ptmaxY}]}], {i, 1, cont1}}];
    T1 = Table[{alea1[[i]], f[alea1[[i]][[1]], alea1[[i]][[2]]]}], {i, 1, cont1}}];
    NunOr = Table[i, {i, 1, cont1}];
    M1 = Select[NunOr, Apply[f, Column[T1, 1][[#]]] <= Min[Column[T1, 2]] &];
    M1 = Min[M1];
    eMes = Flatten[Column[T1, 1][[M1]]; (*Print[eMes]*);
    For[i = 1, And[i <= 52, Abs[Apply[f, {x0, y0}] - Apply[f, eMes]] > 10^(-10),
      Norm[{x0, y0} - eMes] > 10^(-10)], i++,
      alea = Table[Random[MultinormalDistribution[eMes, {{(ptmaxX - ptminX) / 2^i - 2},
        0}, {(ptmaxY - ptminY) / 2^i - 2}]}], {j, 1, cont}}];
      Tes = Table[{alea[[k]], f[alea[[k]][[1]], alea[[k]][[2]]]}], {k, 1, cont}}];
      M2 = Select[Table[i, {i, 1, cont}],
        Apply[f, Column[Tes, 1][[#]]] <= Min[Column[Tes, 2]] &]; M2 = Min[M2];
      eMes = If[Apply[f, Flatten[Column[Tes, 1][[M2]]]] <= Apply[f, eMes],
        Flatten[Column[Tes, 1][[M2]]], eMes];
      Tuba = {Max[Append[Tuba, i]]};
      Luba = {Abs[Apply[f, eMes] - f[x0, y0]] / Abs[f[x0, y0]], Norm[eMes]};
    ]
  ]; Uba = Append[Uba, Tuba[[1]]]; Ruba = Append[Ruba, {Tuba[[1]], Luba}];
]

```

Fig. 3.1. The Mathematica implementation of the CGM algorithm

fully downloadable from the author’s web page). For general purposes, we may extend f to the whole space by defining $f(x) = A$ (A large enough) for $x \notin K$.

The algorithm introduced converges to a global minimizer under the hypothesis of continuity of the function f defined on a compact set. For notational purposes given three random variables X, Y, Z and $(a, b) \in \mathbb{R}^n \times \mathbb{R}^{n \times n}$ we will write $X \in \mathcal{N}(Y, Z)$ to mean that conditionally on $Y = a, Z = b, X \in \mathcal{N}(a, b)$, that is, X has Gaussian distribution with mean a and covariance matrix b .

Theorem 2. *Let $f : K \rightarrow \mathbb{R}$ be a real valued continuous function defined over K , a compact set in \mathbb{R}^n , and let z be an unique global minimizer of f in K , that is: $f(z) = \min_{x \in K} f(x)$ and for all $x \in K$ we have $f(z) < f(x)$.*

For each $N \in \mathbb{N} \setminus \{0\}$ fixed, define almost surely and recursively the sequence $(T_j^N)_{j \in \mathbb{N}}$ by:

$$T_0^N = \left\{ X_{i_0}^0 : f(X_{i_0}^0) = \min_{1 \leq i \leq N} f(X_i^0) \quad X_1^0, \dots, X_N^0 \in \mathcal{U}(K) \text{ i.i.d.} \right\} .$$

Next, for all $j \geq 1$

$$T_{j+1}^N := \begin{cases} T_j^N & \text{if } f(T_j^N) < \min_{1 \leq i \leq N} \left\{ f(X_i^{j+1}) : X_i^{j+1} \in \mathcal{N}(T_j^N, \frac{c}{2^{j+1}}) \text{ i.i.d.} \right\} \\ X_{i_0}^{j+1} & \text{if } f(X_{i_0}^{j+1}) = \min_{1 \leq i \leq N} \left\{ f(X_i^{j+1}) : X_i^{j+1} \in \mathcal{N}(T_j^N, \frac{c}{2^{j+1}}) \text{ i.i.d.} \right\} \\ & \leq f(T_j^N) . \end{cases}$$

Then, for all $N \geq 1$ fixed, the sequence $(T_j^N)_{j \geq 0}$ is a uniformly integrable martingale which converges almost surely to a random variable T^N and the sequence $(T^N)_{N \geq 1}$ converges almost surely to z , the unique global minimizer of f .

Proof. For all $j \geq 1$ define the σ algebras $\mathcal{G}_j^N = \sigma(T_0^N, \dots, T_j^N)$ and the sets:

$$A_{j+1}^N = \left\{ f(T_j^N) < \min_{1 \leq i \leq N} \left\{ f(X_i^{j+1}) : X_i^{j+1} \in \mathcal{N}(T_j^N, \frac{c}{2^{j+1}}) \text{ i.i.d.} \right\} \right\} \subset \Omega .$$

As a first fact, we have obviously that $A_{j+1}^N \in \mathcal{G}_j^N$. Let us remark that $(T_j^N)_{j \geq 0}$ is a martingale with respect to the filtration $(\mathcal{G}_j^N)_{j \geq 0}$. This is a consequence of:

$$\begin{aligned} \mathbb{E}[T_{j+1}^N \mid \mathcal{G}_j^N] &= \mathbb{E}[T_j^N \mathbb{1}_{A_{j+1}^N} + X_{i_0}^{j+1} \mathbb{1}_{(A_{j+1}^N)^c} \mid \mathcal{G}_j^N] = \\ &= T_j^N \mathbb{1}_{A_{j+1}^N} + \mathbb{1}_{(A_{j+1}^N)^c} \mathbb{E}[X_{i_0}^{j+1} \mid \mathcal{G}_j^N] = T_j^N , \end{aligned}$$

as we have, by the definitions,

$$\mathbb{E}[X_{i_0}^{j+1} \mid \mathcal{G}_j^N] = \mathbb{E}[X_{i_0}^{j+1} \mid T_0^N, \dots, T_j^N] = \mathbb{E}[X_{i_0}^{j+1} \mid T_j^N] = T_j^N .$$

As a third fact, we notice that as $T_j^N \in K$, which is a compact set, we then have for some constant $M > 0$ that $\|T_j^N\|_1 \leq M$. As a consequence of these

three facts $(T_j^N)_{j \geq 0}$ is a uniformly integrable martingale which converges almost surely to an integrable random variable T^N . Observe now that, by construction, $f(T_{j+1}^N) \leq f(T_j^N)$ almost surely and so we have that $(f(T_j^N))_{j \geq 0}$ decreases to $f(T^N)$. Let us remark that for all i, j we have $f(T^N) \leq f(X_i^j)$. In fact, by definition:

$$\begin{aligned} \min_{1 \leq i \leq N} f(X_i^{j_0+1}) &\geq \begin{cases} f(T_{j_0}^N) \text{ in } A_{j_0+1}^N \\ f(X_{i_0}^{j_0+1}) \text{ in } (A_{j_0+1}^N)^c \end{cases} = \\ &= f(T_{j_0+1}^N) \mathbb{1}_{A_{j_0+1}^N} + f(T_{j_0+1}^N) \mathbb{1}_{(A_{j_0+1}^N)^c} = f(T_{j_0+1}^N), \end{aligned}$$

so, if for some i_0, j_0 we had

$$f(T^N) > f(X_{i_0}^{j_0+1}) = \min_{1 \leq i \leq N} f(X_i^{j_0+1}) \geq f(T_{j_0+1}^N),$$

we would also have $f(T^N) > f(T_{j_0+1}^N)$, which contradicts the properties of $(f(T_j^N))_{j \geq 0}$. We will now show that the sequence $(f(T^N))_{n \geq 1}$ converges to $f(z)$ in probability. For that purpose, we recall the definition and some simple properties of $E_t := \{x \in K : f(x) < t\}$ the set of points of K having a level, given by f , less than t . First, the monotony: $t < s \Rightarrow E_t \subseteq E_s$; secondly, E_s is open: $x_0 \in E_t \Rightarrow \exists \delta > 0 \ B_{\mathbb{R}^n}(x_0, \delta) \subseteq E_t$; finally, $E_t \neq \emptyset \Rightarrow z \in E_t$. Observe now that for all $\omega \in \Omega$ and all $\eta > 0$:

$$|f(T^N(\omega)) - f(z)| > \eta \Leftrightarrow \begin{cases} f(T^N(\omega)) < f(z) - \eta \text{ which is impossible;} \\ f(T^N(\omega)) > f(z) + \eta \Rightarrow T^N(\omega) \notin E_{f(z)+\eta}. \end{cases}$$

As a consequence, for all i, j we have $X_i^j(\omega) \notin E_{f(z)+\eta}$, as otherwise we would have $f(X_i^j(\omega)) < f(z) + \eta < f(T^N(\omega)) \leq f(X_i^j(\omega))$, which is impossible. As a result, we finally have:

$$\{|f(T^N) - f(z)| > \eta\} \subseteq \bigcap_{j=0}^{+\infty} \bigcap_{i=0}^N \{X_i^j \notin E_{f(z)+\eta}\},$$

which implies that $\mathbb{P}[|f(T^N) - f(z)| > \eta]$ is bounded above, for instance, by:

$$\left(\inf_{0 \leq j < +\infty} \mathbb{P}[X_i^j \notin E_{f(z)+\eta}] \right)^N \leq \mathbb{P}[X_i^0 \notin E_{f(z)+\eta}]^N.$$

Now, X_i^0 being uniformly distributed over K we have, with η small enough: $\mathbb{P}[X_i^0 \notin E_{f(z)+\eta}] = \lambda(E_{f(z)+\eta}^c) / \lambda(K) < 1$. So, we have as wanted, for all $\eta > 0$ small enough: $\lim_{n \rightarrow +\infty} \mathbb{P}[|f(T^N) - f(z)| > \eta] = 0$.

We now observe that the above convergence is, in fact, almost sure convergence, that is, $(f(T^N))_{N \geq 1}$ converges to $f(z)$ almost surely. This is a consequence of the well known fact that a non increasing sequence of random variables converging in probability, converges almost surely and the facts proved above that show:

$$\begin{aligned}
 f(T_j^{N+1}) &\leq f(T_j^N) \\
 \downarrow_{j \rightarrow +\infty} \quad \downarrow_{j \rightarrow +\infty} \\
 f(T^{N+1}) &\leq f(T^N).
 \end{aligned}$$

Consider $\Omega \subset \Omega$ such that $\mathbb{P}[\Omega] = 1$ such that the above convergence takes place over Ω and observe that for every $\omega \in \Omega$ the sequence $(T^N(\omega))_{N \geq 1}$ is a sequence of points in the compact set K . As a consequence, every convergent subsequence $(T^{N_k}(\omega))_{N \geq 1}$ of $(T^N(\omega))_{N \geq 1}$ converges to z . In fact, if $\lim_{k \rightarrow +\infty} T^{N_k}(\omega) = y \in K$ then $\lim_{k \rightarrow +\infty} f(T^{N_k}(\omega)) = f(y)$ and by the result we just proved $\lim_{k \rightarrow +\infty} f(T^{N_k}(\omega)) = f(z)$. This implies that $f(y) = f(z)$ and as z is a unique minimizer we have that $y = z$. We now conclude the proof of the theorem by noticing that $(T^N(\omega))_{N \geq 1}$ converges to z because if otherwise we would have: $\exists \epsilon > 0 \forall N \exists N_m > N \mid T^{N_m}(\omega) - z \mid > \epsilon$. As $(T^{N_m}(\omega))_{m \geq 1}$ is a sequence of points in K which is a compact set, by Bolzano Weierstrass theorem it has a convergent subsequence. This subsequence must converge to z which can not occur by the definition of $(T^{N_m}(\omega))_{m \geq 1}$.

4 Computational Results

CGM algorithm was compared with other algorithms. With Styblinski-Tang function, we compared algorithms A (simple random search), B (localized random search), C (enhanced random search), from [9, p. 38–48], and ARS (accelerated random search) from [1]. The following notations are used. N is the number of random points drawn at each repetition; M will denote the number of repetitions in the simulation; J will be number of steps in the repetition; \bar{J} represents the sample mean of the number of steps J necessary to achieve a prescribed result, taken over the whole set of repetitions of the simulation; $SD(J)$ is the sample standard deviation of J . The stopping criterion for the number of steps j

```

Jota = 400; Uva = {};
Ce = 2^40.5;
Rol = 10^(-4);
For[j = 1, j ≤ Jota, j++,
  XisEne = {Random[Real, {xm, xM}], Random[Real, {xm, xM}]};
  ErEne = 1;
  ind = 0;
  For[
    i = 1,
    And[i ≤ 25000, Abs[Apply[f, {x0, y0}] - Apply[f, XisEne]] > 10^(-10),
    Norm[{x0, y0} - XisEne] > 10^(-10)], i++,
    YupEne = {Random[Real, {XisEne[[1]] - ErEne * (xM - xm) / 2,
      XisEne[[2]] + ErEne * (xM - xm) / 2}], Random[Real,
      {XisEne[[1]] - ErEne * (xM - xm) / 2, XisEne[[1]] + ErEne * (xM - xm) / 2}]};
    If[Apply[f, YupEne] < Apply[f, XisEne], And[XisEne = YupEne, ErEne = 1],
    ErEne = ErEne / Ce];
    If[ErEne < Rol, ErEne = 1,];
    ind = i;
    Vaca = {ind, Abs[Apply[f, XisEne] - f[x0, y0]] / Abs[f[x0, y0]], Norm[XisEne]};
  ]; Uva = Append[Uva, Vaca];
]

```

Fig. 4.1. The Mathematica implementation of the ARS algorithm used

in the simulation will be: $j \leq J_0 \wedge |f(z_n) - f(z)| > 10^{-10} \wedge |z_n - z| > 10^{-10}$, J_0 being the number of steps we decide to impose as an upper bound and z_n being the estimated minimizer at step n of a repetition. A true minimizer of the function is z . The function evaluation accuracy criterion after j steps is:

$$\Delta f(z_j) = \begin{cases} (f(z_j) - f(z))/f(z) & \text{if } f(z) \neq 0 \\ f(z_j) & \text{if } f(z) = 0. \end{cases}$$

The function argument evaluation accuracy criterion after j steps is:

$$\Delta z_j = \begin{cases} \|(z_j - z)/z\| & \text{if } z \neq 0 \\ \|z_j\| & \text{if } z = 0. \end{cases}$$

The averages or sample means over M repetitions with $j(k)$ steps at repetition k are: $\overline{\Delta z} = (1/M) \sum_{k=1}^M \Delta z_{j(k)}$, $\overline{z} = (1/M) \sum_{k=1}^M z_{j(k)}$, $\overline{\Delta f(z)} = (1/M) \sum_{k=1}^M \Delta f(z_{j(k)})$ and $\overline{f(z)} = (1/M) \sum_{k=1}^M f(z_{j(k)})$.

Table 1. Styblinski-Tang function; Repetitions: 400; ARS number of evaluations: 25000

Algorithm	$\overline{f(z)}$	$SD(f(z))$	\overline{E}	$SD(E)$
A	-78.2732	2.8×10^{-3}	25000	-
B	-78.3201	6.1581×10^{-4}	25000	-
C	-78.3201	5.9301×10^{-4}	25000	-
ARS (381)	-78.3323	1.35255×10^{-8}	24778	1818
CGM	-78.3323	2.72116×10^{-11}	15783	970

Table 2. CGM algorithm; random draws: 500; maximum number of steps: 50

Function	\overline{J}	$SD(J)$	$\overline{\Delta f(z)}$	$\overline{\Delta z}$
Gaussian 1 (371)	33.372	1.91627	9.34019×10^{-12}	8.96686×10^{-7}
Gaussian 2 (97)	37.0103	1.70474	7.38679×10^{-12}	2.49709×10^{-7}
Griewank	30.475	1.873	4.65529×10^{-11}	9.63914×10^{-6}
Himmelblau	32.1675	1.91012	4.21595×10^{-11}	1.09551
Jennrich-Sampson	49.7375	2.13541	4.90973×10^{-11}	8.77548×10^{-7}
Rastrigin (399)	32.3885	1.75567	2.29124×10^{-11}	4.93828×10^{-7}
Rosenbrock	30.5325	1.90868	4.30236×10^{-11}	6.17981×10^{-6}
Freudenstein-Roth	33.9725	1.83033	4.42997×10^{-11}	5.23501×10^{-7}
Styblinski-Tang	31.5675	1.93903	5.13851×10^{-13}	3.45016×10^{-7}

Table 3. ARS algorithm; Maximum number of function evaluations: 25000

Function	\overline{N}	$SD(N)$	$\overline{\Delta f(z)}$	$\overline{\Delta z}$
Gaussian 1	24984.8	303.95	5.34489×10^{-9}	2.03585×10^{-5}
Gaussian 2 (356)	25000	0	7.45836×10^{-8}	2.27671×10^{-5}
Griewank	25000	0	4.75247×10^{-8}	2.90102×10^{-4}
Himmelblau	25000	0	5.03447×10^{-1}	6.76862×10^{-1}
Jennrich-Sampson	24900.3	1426.31	1.00239×10^{-9}	1.24142×10^{-5}
Rastrigin	23991.4	3954.14	7.65517×10^{-10}	2.67892×10^{-6}
Rosenbrock (398)	24518.6	2543.41	2.89627×10^{-9}	4.59241×10^{-5}
Freudenstein-Roth	25000	0	4.80997×10^{-2}	1.63671×10^{-2}
Styblinski-Tang (382)	24777.9	1862.48	7.08203×10^{-11}	3.6433×10^{-6}

ARS and the CGM algorithms were compared for nine different test functions from [1] and [4]. The numerical results presented in the tables show that CGM outperforms all other algorithms tried, in precision and with less function evaluations. For one of the test functions (Gaussian 2) a further test, run with 2700 random draws at each step of the repetitions (instead of 500 used for the tables study) gave a result with the prescribed precision for all repetitions, in accordance with theorem 2. In the tables, the number between parenthesis in a given line report correct locations of the global minimum among the repetitions performed.

5 Conclusion

To achieve its goal, any random search algorithm has simultaneously to detect the region where the global minimum is located and to achieve enough precision in the calculation of the minimizer. Practically, and to the limits of machine and software precision used this is obtained, respectively, by an increasing number of random trials and, by concentrating these trials in the favorable region. CGM algorithm, hereby introduced and studied, always attained better precision than ARS; we got also perfect global minimum location for all the test functions tried, provided (in three cases) the number of random draws was sufficiently augmented. The CGM convergence result was proved under mild but fully verifiable hypothesis in sharp contrast with our formulation of a Solis and Wets's type theorem for adaptive random search with an hypothesis of difficult if not impossible verification even, in very simple cases.

References

1. Appel, M. J., Labarre, R., Radulović, D.: On accelerated random search. *SIAM J. Optim.*, **14**, (2003) no. 3, 708–731.
2. Guimier, A.: Modélisation d'algorithmes d'optimisation à stratégie aléatoire. *Calcolo*. **23** (1986) no. 1, 21–34.
3. Horst, R., Pardalos, P. M. (ed.): *Handbook of Global Optimization*. Kluwer Academic Publishers 1995.
4. Moré, J. J., Garbow, B. S., Hillstom, K. E.: Testing Unconstrained Optimization Software, *ACM Transactions on Mathematical Software* **7**, No.1, (1981) 17–41.
5. Pintér, J. D.: *Global optimization in action. Continuous and Lipschitz optimization: algorithms, implementations and applications.*, Nonconvex Optimization and Its Applications. 6. Dordrecht: Kluwer Academic Publishers. xxvii 1996.
6. Raphael, B., Smith, I.F.C.: A direct stochastic algorithm for global search. *Appl. Math. Comput.* **146**, No.2-3, (2003) 729–758.
7. Shi, D., Peng, J.: A new theoretical framework for analyzing stochastic global optimization algorithms. *J. Shanghai Univ.* **3**, No.3, (1999) 175–180.
8. Solis, F. J., Wets, R. J.-B.: Minimization by random search techniques. *Math. Oper. Res.*, **6**, (1981) no. 1, 19–30.
9. Spall, J. C.: *Introduction to stochastic search and optimization. Estimation, simulation, and control*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Hoboken, NJ: Wiley. xx, 2003.

10. Stephens, C.P., Baritomba, W.: Global optimization requires global information. *J. Optimization Theory Appl.* **96**, (1998) No.3, 575–588.
11. Yin, G.: Rates of convergence for a class of global stochastic optimization algorithms. *SIAM J. Optim.* **10**, No.10, (1999) 99–120 .
12. Zabinsky, Z. B.: *Stochastic adaptive search for global optimization*. Nonconvex Optimization and Its Applications 72. Boston, MA: Kluwer Academic Publishers. xviii, 2003.

Finding the Number of Clusters Minimizing Energy Consumption of Wireless Sensor Networks*

Hyunsoo Kim and Hee Yong Youn

School of Information and Communications Engineering,
Sungkyunkwan University, 440-746, Suwon, Korea
bayes1@hanmail.net, youn@ece.skku.ac.kr

Abstract. Wireless sensor network consisting of a large number of small sensors with low-power can be an effective tool for the collection and integration of data in a variety of environments. Here data exchange between the sensors and base station need to be designed to conserve the limited energy resources of the sensors. Grouping the sensors into clusters has been recognized as an efficient approach for saving the energy of the sensor nodes. In this paper we propose an analytical model based on homogenous spatial Poisson process, which allows the number of clusters in a sensor network minimizing the total energy spent for data exchange. Computer simulation on various size sensor networks with the LEACH algorithm reveals that the proposed model is very accurate. We also compare the proposed model with an earlier one, and it turns out that the proposed model is more accurate.

Keywords: Energy consumption, LEACH, number of clusters, Poisson process, wireless sensor network.

1 Introduction

Recent developments in wireless sensor network have motivated the growth of extremely small, low-cost sensors that possess the capability of sensing, signal processing, and wireless communication. The wireless sensor network can be expanded at a cost much lower than conventional wired sensor network. Here each sensor is capable of detecting the condition around it such as temperature, sound, and the presence of other objects. Recently, the design of sensor networks has gained increasing importance due to their potential for civil and military applications such as combat field surveillance, security, and disaster management. The smart dust project at the University of California, Berkeley [7, 8, 12] and WINS project at UCLA [9] are attempting to build extremely small sensors of low-cost allowing autonomous sensing and communication in a cubic millimeter

* This research was supported in part by the Ubiquitous Autonomic Computing and Network Project, 21st Century Frontier R&D Program in Korea and the Brain Korea 21 Project in 2005. Corresponding author: Hee Yong Youn.

range. The system processes the data gathered from the sensors to monitor the events in an area of interest.

Prolonged network lifetime, scalability, and load balancing are important requirements for many wireless sensor network applications. A longer lifetime of sensor networks can be achieved through optimized applications, operating systems, and communication protocols. If the distance between the nodes is small, energy consumption is also small. With the direct communication protocol in a wireless sensor network, each sensor sends its data directly to the base station. When the base station is far away from the sensors, the direct communication protocol will cause each sensor to spend a large amount of power for data transmission [5]. This will quickly drain the battery of the sensors and reduce the network lifetime. With the minimum energy routing protocol, the sensors route data ultimately to the base station through intermediate sensors. Here only the energy of the transmitter is considered while energy consumption of the receivers is neglected in determining the route. The low-energy adaptive clustering hierarchy (LEACH) scheme [9] includes the use of energy-conserving hardware. Here sensors in a cluster detect events and then transmit the collected information to the cluster head. Power consumption required for transmitting data inside a cluster is lower than with other routing protocols such as minimum transmission energy routing protocol and direct communication protocol due to smaller distances to the cluster head than to the base station.

There exist various approaches for the evaluation of the routing protocols employed in wireless sensor networks. In this paper our main interest is energy consumption of the sensors in the network. Gupta and Kumar [2,3] have analyzed the capacity of wireless ad-hoc networks and derived the critical power with which a node in a wireless ad-hoc network communicates to form a connected network with probability one. A sensor in a wireless sensor network can directly communicate with only the sensors within its radio range. To enable communication between the sensors not within each other's communication range, the sensors need to form a cluster with the neighboring sensors. An essential task in sensor clustering is to select a set of cluster heads among the sensors in the network, and cluster the rest of the sensors with the cluster heads. The cluster heads are responsible for coordination among the sensors within their clusters, and communication with the base station. [4] have suggested an approach for cluster formation which ensures the expected number of clusters in LEACH. A model deriving the number of clusters with which the energy required for communication is minimized was developed in [10]. Energy consumption required for communication depends on the distance between the transmitter and receiver. Here, the expected distance between the cluster head and non-cluster head was computed. The distance from a sensor to the base station was not computed but fixed. [11] proposed a method for finding the probability for a sensor to become a cluster head. They directly computed the expected distance from a sensor to the base station with uniform distribution in a bounded region. They assumed that the distance between a cluster head and non-cluster heads in a cluster de-

depends only on the density of the sensors distributed with a homogeneous Poisson process in a bounded region. More details are given in Section 2.3.

In this paper we develop a model finding the number of cluster heads which allows minimum energy consumption of the entire network using homogeneous spatial Poisson process. [11] modeled the expected distance between a cluster head and other sensors in deciding the number of cluster heads. Here, we additionally include the radio range of the cluster head and the distribution of energy dissipation of the sensors for more accurate prediction of the number. With the proposed model the network can determine, a priori, the number of cluster heads in a region of distributed sensors. The validity of the proposed model is verified by computer simulation, where the clustering is implemented using the LEACH algorithm. It reveals that the number of clusters obtained by the developed model is very close to that of the simulation with which the energy consumption of the network is minimum. We also compare the proposed model with [11], and it consistently produces more accurate value than [11].

The rest of the paper is presented as follows. In Section 2 we review the concept of LEACH, sensor clustering, and previous approaches for modeling the number of clusters. Section 3 presents the proposed model finding the number of clusters which minimizes energy consumption. Section 4 demonstrates the effectiveness of the proposed model by computer simulation and comparison with the earlier one. We provide a conclusion in Section 5.

2 Preliminaries

2.1 The LEACH Protocol

LEACH is a self organizing, adaptive clustering protocol that employs a randomization approach to evenly distribute energy consumption among the sensors in the network. In the LEACH protocol, the sensors organize themselves into clusters with one node acting as the local base station or cluster head. If the cluster heads are fixed throughout the network lifetime as in the conventional clustering algorithms, the selected cluster heads would die quickly, ending the useful lifetime of all other nodes belonging to the clusters. In order to circumvent the problem, the LEACH protocol employs randomized rotation of the cluster head such that all sensors have equal probability to be a cluster head in order not to spent the energy of only specific sensors. In addition, it carries out local data fusion to compress the data sent from the cluster heads to the base station. This can reduce energy consumption and increase sensor lifetime. Clusters can be formed based on various properties such as communication range, number and type of sensors, and geographical location.

2.2 Sensor Clustering

Assume that each sensor in the wireless sensor network becomes a cluster head with a probability, p . Each sensor advertises itself as a cluster head to the sensors within its radio range. We call the cluster heads the voluntary cluster heads. For

the sensors not within the radio range of the cluster head, the advertisement is forwarded. A sensor receiving an advertisement which does not declare itself as a cluster head joins the cluster of the closest cluster head. A sensor that is neither a cluster head nor has joined any cluster becomes a cluster head; we call the cluster heads the forced cluster heads. Since forwarding of advertisement is limited by the radio range, some sensors may not receive any advertisement within some time duration, t , where t is the time required for data to reach the cluster head from a sensor within the radio range. It then can infer that it is not within radio range of any voluntary cluster head, and hence it becomes a forced cluster head. Moreover, since all the sensors within a cluster are within radio range of the cluster head, the cluster head can transmit the collected information to the base station after every t units of time. This limit on radio range allows the cluster heads to schedule their transmissions. Note that this is a distributed algorithm and does not require clock synchronization between the sensors.

The total energy consumption required for the information gathered by the sensors to reach the base station will depend on p and radio range, r . In clustering the sensors, we need to find the value of p that would ensure minimum energy consumption. The basic idea of the derivation of the optimal value of p is to define a function of energy consumption required to transmit data to the base station, and then find the p value minimizing it. The model needs the following assumptions.

- The sensors in the wireless sensor network are distributed as per a homogeneous spatial Poisson process of intensity λ in 2-dimensional space.
- All sensors transmit data with the same power level, and hence have the same radio range, r .
- Data exchanged between two sensors not within each others' radio range is forwarded by other sensors.
- Each sensor uses 1 unit of energy to transmit or receive 1 unit of data.
- A routing infrastructure is in place; hence, when a sensor transmits data to another sensor, only the sensors on the routing path forward the data.
- The communication environment is contention and error free; hence, sensors do not have to retransmit any data.

2.3 Previous Modeling

Energy consumption required for communication depends on the distance between the transmitter and receiver, i.e., the distance between the sensors. In [11], first, the expected distance from a sensor to the base station is computed. Then the expected distance from the non-cluster head to the cluster head in a Voronoi cell (a cluster) is obtained. Assume that a sensor becomes a cluster head with probability p . Here the cluster-heads and the non-cluster heads are distributed as independent homogeneous spatial Poisson processes of intensity $\lambda_1 = p\lambda$ and $\lambda_0 = (1 - p)\lambda$, respectively. Each non-cluster head joins the cluster of the closest cluster-head to form a Voronoi tessellation [6]. The plane is divided into zones called the Voronoi cells, where each cell corresponds to a Poisson process with

intensity λ_1 , called nucleus of Voronoi cells. Let N_v be the random variable denoting the number of non-cluster heads in each Voronoi cell, and N be the total number of sensors in a bounded region. If L_v is the total length of all segments connecting the non-cluster heads to the cluster head in a Voronoi cell, then

$$E[N_v|N = n] = \lambda_0/\lambda_1, \quad E[L_v|N = n] = \lambda_0/2\lambda_1^{3/2}. \quad (1)$$

Note that these expectations depend only on the intensities λ_0 and λ_1 of a Voronoi cell. Using the expectations of the lengths above, an optimal value p minimizing the total energy spent by the network is found.

3 The Proposed Scheme

3.1 Motivation

For a wireless sensor network of a large number of energy constrained sensors, it is important to fastly group the sensors into clusters to minimize the energy used for communication. Note that energy consumption is directly proportional to the distance between the cluster heads and non-cluster heads and the distance between the cluster head and base station. We thus focus on the functions of the distances for minimizing the energy spent in the wireless sensor network. Note that the earlier model of Equation (1) depends only on the intensity of the sensor distribution. In this paper, however, the expected distance between the sensors and cluster head is modeled by including p , and the number of sensors, N , and the size of the region for obtaining more accurate estimation on the number of clusters. The number is decided by the probability, p , which minimizes the energy consumed to exchange the data. We next present the proposed model.

3.2 The Proposed Model

In the sensor network the expected distance between the cluster heads to the base station and the expected distance between the sensors to the cluster head in a cluster depend on the number of sensors, the number of clusters, and the size of the region. The expected distance between a sensor and its cluster head decreases while that between a cluster head and the base station increases as the number of clusters increases in a bounded region. An opposite phenomenon is observed when the number of clusters decreases. Therefore, an optimal value of p in terms of energy efficiency needs to be decided by properly taking account the tradeoff between the communication overhead of sensor-to-cluster head and cluster head-to-base station. Figure 1 illustrates this aspect where Figure 1(a) and (b) has 14 clusters and 4 clusters, respectively. Notice that the clusters in Figure 1(a) have relatively sparse links inside the clusters compared to those in Figure 1(b), while there exist more links to the base station.

Let S denote a bounded region of a plane and $X(S)$ does the number of sensors contained in S . Then $X(S)$ is a homogeneous spatial Poisson process if it distributes the Poisson postulates, yielding a probability distribution

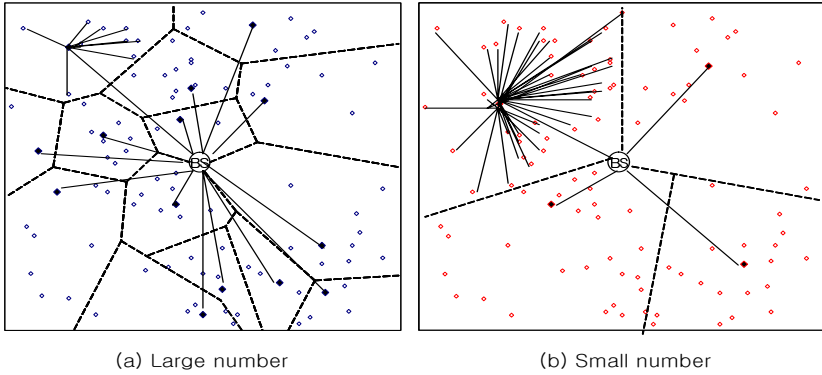


Fig. 1. Comparison of the structures having a large and small number of clusters

$$P\{X(S) = n\} = \frac{[\lambda A(S)]^n e^{-\lambda A(S)}}{n!}, \quad \text{for } A(S) \geq 0, \quad n = 0, 1, 2, \dots$$

Here λ is a positive constant called the intensity parameter of the process and $A(S)$ represents the area of region S .

If region S is a square of side length, M , then the number of sensors in it follows a Poisson distribution with a mean of $\lambda A(S)$, where $A(S)$ is M^2 . Assume that there exist N sensors in the region for a particular realization of the process. If the probability of becoming a cluster head is, p , then Np sensors will become cluster heads on average. Let $D_B(x, y)$ be a random variable denoting the distance between a sensor located at (x, y) and the base station. Let P_S be the probability of existence of sensors uniformly distributed in region S . Without loss of generality, we assume that the base station is located at the center of the square region (i.e. the origin coordinate). Then, the expected distance from the base station to the sensors is given by

$$\begin{aligned} E[D_B(x, y)|X(S) = N] &= \int \int_S D_B(x, y) \cdot P_S \, dS \\ &= \int_{-M/2}^{M/2} \int_{-M/2}^{M/2} \sqrt{x^2 + y^2} \frac{1}{M^2} \, dx \, dy \\ &= 0.3825M \end{aligned} \tag{2}$$

Since there exist Np cluster heads on average and the location of a cluster head is independent of those of other cluster heads, the total length of the segments from all the cluster heads to the base station is $0.3825NpM$.

Since a sensor becomes a cluster head with a probability p , we expect that the cluster head and other sensors are distributed in a cluster as an independent homogeneous spatial Poisson process. Each sensor joins the cluster of the closest cluster head to form a cluster. Let $X(C)$ be the random variable denoting the number of sensors except the cluster head in a cluster. Here, C is the area of a cluster. Let D_C be the distance between a sensor and the cluster head in a

cluster. Then, according to the results of [1], the expected number of non-cluster heads in a cluster and the expected distance from a sensor to the cluster head (assumed to be at the center of mass of the cluster) in a cluster are given by

$$E[X(C)|X(S) = N] = \frac{1}{p} - 1, \tag{3}$$

$$\begin{aligned} E[D_C|X(S) = N] &= \int \int_C \sqrt{x^2 + y^2} k(x, y) \, dA(C) \\ &= \int_0^{2\pi} \int_0^{M/\sqrt{Np\pi}} r^2 \frac{Np}{M^2} \, dr d\theta = \frac{2M}{3\sqrt{Np\pi}}, \end{aligned} \tag{4}$$

respectively. Here, region C is a circle with radius $M/\sqrt{Np\pi}$. The sensor density of the cluster, $k(x, y)$, is uniform, and it is approximately M^2/Np .

Let E_C be the expected total energy used by the sensors in a cluster to transmit one unit of data to their respective cluster head. Since there are Np clusters, the expected value of E_C conditioned on $X(S) = N$ is given by,

$$\begin{aligned} E[E_C|X(S) = N] &= N(1 - p) \cdot \frac{E[D_C|X(S) = N]}{r} \\ &= N^{\frac{1}{2}} \cdot \frac{2M}{3r\sqrt{\pi}} \frac{1 - p}{\sqrt{p}}. \end{aligned} \tag{5}$$

If the total energy spent by the cluster heads to transmit the aggregated information to the base station is denoted by E_B , then

$$\begin{aligned} E[E_B|X(S) = N] &= Np \cdot \frac{E[D_B|X(S) = N]}{r} \\ &= \frac{0.3825NpM}{r}. \end{aligned} \tag{6}$$

Let E_T be the total energy consumption with the condition of $X(S) = N$ in the network. Then

$$E[E_T|X(S) = N] = E[E_C|X(S) = N] + E[E_B|X(S) = N]. \tag{7}$$

Taking the expectation of Equation (7), the total energy consumption of the network is obtained.

$$\begin{aligned} E[E_T] &= E[E[E_T|X(S) = N]] \\ &= E[X(S)^{1/2}] \cdot \frac{2M}{3r\sqrt{\pi}} \frac{1 - p}{\sqrt{p}} + E[X(S)] \cdot \frac{0.3825pM}{r}, \end{aligned} \tag{8}$$

where $E[\cdot]$ is expectation of a homogeneous Poisson process.

$E[E_T]$ will have a minimum value for a value of p , which is obtained by the first derivative of Equation (8);

$$2c_2p^{3/2} - c_1(p + 1) = 0, \tag{9}$$

where $c_1 = 2M \cdot E[(X(S))^{1/2}]/3\sqrt{\pi}$ and $c_2 = 0.3825M \cdot E[X(S)]$.

Equation (9) has three roots, two of which are imaginary. The second derivative of Equation (8) is positive and log concave for the only real root of Equation (9), and hence the real root minimizes the total energy consumption, $E[E_T]$.

The only real root of Equation (9) is as follows.

$$p = \frac{0.0833c_1^2}{c_2^2} + \frac{0.1050(c_1^4 + 24c_1^2c_2^2)}{c_2^2(2c_1^6 + 72c_1^4c_2^2 + 432c_1^2c_2^4 + 83.1384c_1^2\sqrt{c_1^2c_2^6 + 27c_2^8})^{1/3}} + \frac{0.0661}{c_2^2}(2c_1^6 + 72c_1^4c_2^2 + 432c_1^2c_2^4 + 83.1384c_1^2\sqrt{c_1^2c_2^6 + 27c_2^8})^{1/3}. \quad (10)$$

4 Performance Evaluation

In order to validate the proposed model, we simulate a network of sensors distributed as a homogeneous Poisson process with various spatial densities in a region. We employ the LEACH algorithm to generate a cluster hierarchy and find how much energy the network spends with the p value obtained using the developed model. For various intensities $\lambda = (0.01, 0.03, 0.05)$ of sensors in a bounded region, we first compute the probability for a sensor to be a cluster head by Equation (9). Table 1 lists the probability obtained using the developed model and the corresponding energy consumption of Equation (8) for different intensities and radio range of a sensor, r . Here, M is set to 100, and thus there exist 100, 300, and 500 sensors if $\lambda = 0.01, 0.03,$ and $0.05,$ respectively.

Table 1. The p probability and corresponding energy consumption with the proposed model

λ	p	E_T		
		$r = 1$	$r = 2$	$r = 3$
0.01	0.147	1398.03	699.02	466.01
0.03	0.099	3000.77	1500.39	1000.26
0.05	0.083	4263.68	2131.84	1421.23

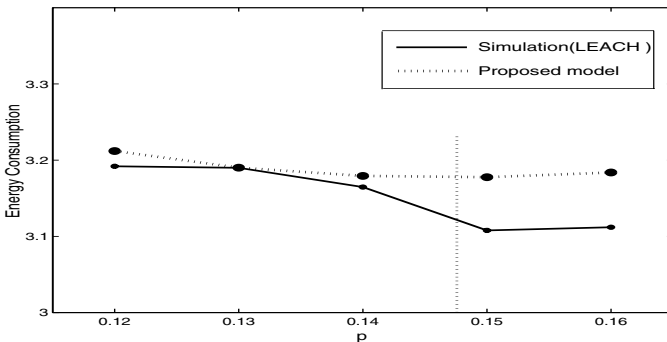


Fig. 2. The total energy consumptions from the proposed model and simulation

Table 2. The p probability and corresponding energy consumption with [11]

λ	p	E_T		
		$r = 1$	$r = 2$	$r = 3$
0.01	0.181	1674.76	837.381	558.25
0.03	0.121	3615.65	1807.83	1205.32
0.05	0.101	5147.63	2573.81	1715.88

Table 3. Comparison of the proposed model with an earlier model [11]

Model	No. of Nodes		
	100	300	500
The proposed model	0.057324	0.1092644	0.299371
[11]	0.059178	0.197071	0.310891
Percentage of energy saving	3.133	2.247	3.706

Note that the validity of the p value estimated by the proposed analytical model can be verified only through actual implementation of the clustering. Figure 2 shows the energy consumed by the entire network as p changes, obtained using the LEACH algorithm which is one of the most popular clustering algorithms for sensor network, when $\lambda = 0.01$ and $r = 1$. Recall that the p value predicted by the proposed model in this case is 0.147, while the optimal value of p obtained from the simulation turns out to be 0.145. The proposed model thus can be said to be very accurate for deciding the p value. We obtain similar results as this for different cases of λ and r values.

We also compare the proposed model and that of [11]. In order to show the relative effectiveness of the proposed model, the probability and the corresponding energy consumption of [11] are obtained and listed in Table 2. Observe from the tables that the probabilities of the proposed model are smaller than [11]. Table 3 summarizes the results of comparison. Here the p values decided from the proposed model and [11] are applied to the LEACH algorithm for obtaining the energy consumed by the network. Notice that the proposed model consistently provides better p values than [11] regardless of the sensor density.

5 Conclusion

The sensors becoming the cluster heads spend relatively more energy than other sensors because they have to receive information from other sensors within their cluster, aggregate the information, and then transmit them to the base station. Hence, they run out of their energy faster than other sensors. To solve this problem, the re-clustering needs to be done periodically or the cluster heads trigger re-clustering when their energy level falls below a certain threshold. We have developed a model deciding the probability a sensor becomes a cluster head,

which minimizes the energy spent by the network for communication. Here the sensors are distributed in a bounded region with homogeneous spatial Poisson process. The validity of the proposed model was verified by computer simulation, where the clustering is implemented using the LEACH algorithm. It revealed that the number of clusters obtained by the developed model is very close to that of the simulation with which the energy consumption of the network is minimum. We also compared the proposed model with [11], and it consistently produces more accurate value than [11].

Here we assumed that the base station is located at the center of the bounded region. We will develop another model where the location of base station can be arbitrary. We will also expand the model for including other factors such as the shape of the bounded region.

References

1. S.G. Foss and S.A. Zuyev, "On a Voronoi Aggregative Process Related to a Bivariate Poisson Process," *Advances in Applied Probability*, Vol. 28, no. 4, 965-981, 1996.
2. P. Gupta and P.R. Kumar, "The capacity of wireless networks," *IEEE Transaction on Information Theory*, Vol. IT-46, No. 2, 388-404, March, 2000.
3. P. Gupta and P.R. Kumar, "Critical power for asymptotic connectivity in wireless networks," *Stochastic Analysis, Control, Optimization and Applications: A Volume in Honor of W. H. Fleming*, 547-566, 1998.
4. W.B. Heinzelman, A.P. Chandrakasan and H. Balakrishnan, "An Application-Specific Protocol Architecture for Wireless Microsensor Networks", *IEEE Trans. on Wireless Communications*, Vol. 1, No. 4, 660-670, 2002.
5. W.R. Heinzelman, A. Cahandrakasan and H. Balakrishnan, "Energy-efficient communication protocol for wireless sensor networks, in the Proceedings of the 33rd Hawaii International Conference on System Sciences, Hawaii, 2000.
6. T. Meng and R. Volkan, "Distributed Network Protocols for Wireless Communication", *In Proc. IEEE ISCAS*, 1998.
7. V. Hsu, J.M. Kahn, and K.S.J. Pister, "Wireless Communications for Smart Dust", *Electronics Research Laboratory Technical Memorandum M98/2*, Feb. 1998.
8. J.M. Kahn, R.H. Katz and K.S.J. Pister, "Next Century Challenges: Mobile Networking for Smart Dust," in the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom 99), 271-278, Aug. 1999.
9. G.J. Pottie and W.J. Kaiser, "Wireless integrated network sensors," *Communications of the ACM*, Vol 43, No. 5, 51-58, May, 2000.
10. T. Rappaport, "Wireless Communications: Principles & Practice", *Englewood Cliffs, NJ: Prentice-Hall*, 1996.
11. S. Bandyopadhyay and E.J. Coyle, "An energy efficient hierarchical clustering algorithm for wireless sensor networks," *IEEE INFOCOM 2003 - The Conference on Computer Communications*, vol. 22, no. 1, 1713-1723, March 2003.
12. B. Warneke, M. Last, B. Liebowitz, Kristofer and S.J. Pister, "Smart Dust: Communication with a Cubic-Millimeter Computer," *Computer Magazine*, Vol. 34, No 1, 44-51, Jan. 2001.

A Two-Echelon Deteriorating Production-Inventory Newsboy Model with Imperfect Production Process

Hui-Ming Wee and Chun-Jen Chung

Department of Industrial Engineering, Chung Yuan Christian University,
Chungli 32023, Taiwan, ROC
weehm@cycu.edu.tw

Abstract. This paper discusses a two-echelon distribution-free deteriorating production-inventory newsboy model. Distinct from previous models, this study considers a deteriorating commodity with the imperfect production process. The model also considers JIT (Just In Time) deliveries and integrates the marketing and manufacturing channels. This model can be used to solve the production and replenishment problem for the manufacturer and the retailer. We derive the optimal production size, the wholesale price and the optimal number of deliveries. The model takes an insight to the defect-rate reduction. We find an upper bound for the optimal number of material's JIT deliveries. The necessary and sufficient conditions are explored to gain managerial insights and economic implications.

1 Introduction

The newsboy problem is very well suited for a single period uncertain demand pattern. This type of problem has significant theoretical and practical implications. Some of the newsboy problem examples are in the inventory management of fashion goods, seasonal sports and apparel industry (Gallego and Moon, [1]). If the manufacturer's wholesale price is too high, the retailer's orders may diminish and the profit may decrease. Our study focuses on determining the optimal wholesale price and the production size policies when the retailer has decided on the order size by the distribution-free newsboy model.

There are several newsboy papers exploring the return problems of the integrated inventory model for the manufacturers and the retailers. Lau and Lau ([2], [3]) developed the pricing and return policies for a single-period item model and studied how the uncertain retail-market demand would affect the expected manufacturer's and the retailer's profits. Lariviere and Porteus [4] incorporated market size and market growth into the model to investigate how market size and market growth affect profits. Shang and Song [5] developed a simple heuristic to minimize 2N newsvendor-type cost functions. When the demand distribution is unknown, the newsboy problem may be classified as distribution-free newsboy problem. Several authors have analyzed the

distribution-free newsboy problem, which the demand distribution is specified by the mean μ and variance σ^2 are. Gallego and Moon [1] analyzed the cases with random yields, fixed ordering cost, and constrained multiple products. Moon and Choi [6] extended the model of Gallego and Moon to consider the balking customers. Moon and Silver [7] studied the distribution-free models of a multi-item newsboy problem with a budget constraint and fixed ordering costs.

However, the above did not consider deterioration in the “manufacturing channel”. The phenomena of deterioration are prevalent and should not be neglected in the model development. Several researchers have studied deteriorating inventory in the past. Ghare and Schrader [8] were the first authors to consider on-going deterioration of inventory. Other authors such as Covert and Philip [9] and Raafat et al. [10] assumed different production policy with different assumptions on the patterns of deterioration. Wee [11] dealt with an economic production quantity model for deteriorating items with partial backordering. Wee and Jong [12] investigated the JIT delivery policy in an integrated production-inventory model for deteriorating items. Yang and Wee [13] provided a review of an integrated deteriorating inventory model.

Quality management, JIT purchasing and manufacturing strategy are widely studied in manufacturing issues. Nassimbeni [14] investigated the relationship between the JIT purchasing and JIT practices. The study demonstrates that purchasing is related to three factors: delivery synchronization, interaction on quality and interaction on design. The buyer or procurer perceived that the quality is a responsibility of the supplier (Nassimbeni, 1995). The JIT purchasing of materials from outside suppliers is notable development of JIT concept. The JIT purchasing ensures that disruption of production can be eradicated immediately. To avoid the loss of production and reduce the inventory level and scrap, the materials received by the suppliers should be of good quality (Reese and Geisel, [15]).

The non-conforming item incurs a greater post-sale service cost than a conforming item. The warranty cost which is included in the post-sale service cost influences the manufacturer’s pricing policy. Lee and Rosenblatt [16] considered the imperfect production process in a traditional EPQ model and derived the optimal number of inspection and production cycle time. Wang and Sheu [17] investigated the imperfect production model with free warranty for the discrete item. They derived an optimal production lot size to minimize the total cost. Wang [18] studied the production process problem to consider the sold products with free warranty. Yeh et al. [19] developed a production inventory model considering free warranty, and derived an optimal production cycle time. To the best of our knowledge, no research on deteriorating two-echelon distribution-free inventory model with imperfect production has been done.

2 Assumptions and Notation

As shown is Figure 1, there are M types of raw materials in the manufacturing process. The retailer must sell the finished products before the expiration date.

After the expiration date the unsold stock has a salvage value. Other assumptions of the seasonal products in the production inventory model development are as follows:

- (a) For a fixed wholesale price, salvage cost, and average demand, the retailer uses the newsboy rule to determine the ordering quantity.
- (b) The manufacturer sums up the orders from retailer as the production lot size.
- (c) The material inventory is controlled by periodic review system, and the backorder is not allowed.
- (d) The lead-time for raw materials is constant. The transportation time is assumed to be zero.
- (e) The production rate is constant and greater than the demand rate.
- (f) Deteriorating rate is constant, and the deteriorated items are not replaced.
- (g) Deterioration of the units is considered only after they have been received into the inventory.
- (h) The manufacturer and the retailer have complete information of each other.
- (i) The replenishment is instantaneous. After a complete production lot, then a delivery batch is dispatched to the retailer.
- (j) At the start of each production cycle, the production process is in an in-control state. After a period of production, the production process may shift from in-control to out-of-control state. The elapsed time is exponentially distributed with a finite mean and variance.
- (k) When the production process shifts to an out-of-control state, the shift cannot be detected till the end of each production period. The process is brought back to an in-control state at the beginning of a new production cycle.
- (l) There are M types materials with JIT multiple deliveries. Type i material is independent of type j material, $i \neq j$.
- (m) The reduction of the imperfect product is proportional to the number of the material deliveries.

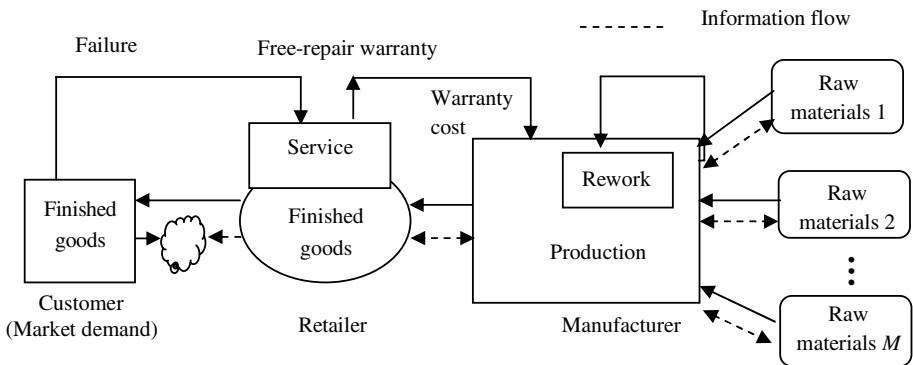


Fig. 1. The behavior of the production-inventory chain

- (n) The selling price is independent of the number of the material deliveries.
- (o) The percentages of the non-conforming items during in-control and out-of control states are assumed to be fixed.

The notation in the model is defined as follows:

Table 1. Notation for model development

μ_c	The mean demand	ϑ_1	Percentage of non-conforming items produced during the production process is in control
σ	The standard deviation	ϑ_2	Percentage of non-conforming items produced during the production process is out of control
$S=(1+a)C_p$	Selling price of the retailer	C_R	Unit rework cost (including inspection cost)
$L_S=b C_p$	Lost sale	C_w	Unit warranty cost
$V=(1-c)C_p$	Salvage	G_m	Unit target profit
$U=dC_p$	Unit transportation cost	u	Unit production cost
$A_1=eC_p$	Unit ordering cost	H	Manufacturer's holding cost
C_S	Production setup cost	L_j	The difference in leadtime for material j (= maximum leadtime -average leadtime)
P	Capacity	C_{mj}	Material's ordering cost of material j
θ	Deteriorating rate	h_{dmj}	The extra handling cost of material j
K	Warranty period	C_{rj}	Unit item cost of material j
A_0	Fixed ordering cost	H_j	Material's holding cost
F	Fixed transportation cost	α_j	Amount of material j needed for unit product
M	The No. of material's type	r_j	The defect-rate-reduction ratio for material j

3 Modeling Description

In this study, the production-inventory model has two phases; phase 1 involves the retailer's policy when the demand is of unknown distribution. The retailer's ordering cost is assumed to be proportion to the ordering quantities. The relevant transportation cost is also incorporated in the model. In phase 2, the manufacturer's production policy is explored. We assume that the manufacturer provides the free-repair replacement for the imperfect product. Based on the result in phase 1, we derive the optimal solution of the manufacturer's production policy.

In practice, demand is influenced by the wholesale price. In addition to the issue of the wholesale price, the demand distribution of a single-period seasonal product is usually unknown. We use the method of distribution-free to analyze the problem. The demand distribution is assumed to be in the worst possible distribution, \mathfrak{S} , and the unknown distribution is G . Let D denote the random demand, $G \in \mathfrak{S}$, with mean μ and variance σ^2 .

To derive the optimal ordering quantities, we define the following relationship:

$$S = (1 + a)C_p; V = (1 - c)C_p; L_s = bC_p; U = dC_p; A_1 = eC_p; \\ 0 < a < 1, 0 < b < 1, 0 < c < 1, 0 < d < 1, 0 < e < 1$$

The retailer determines the ordering quantity using the newsboy rule; the ordering quantity should satisfy the following formula:

$$ER^G = S \cdot E(\min\{Q, D\}) + V \cdot E(Q - D)^+ \tag{1a}$$

$$EC^G = A_0 + F + (A_1 + C_p + U)Q + L_s E(D - Q)^+ \tag{1b}$$

$$EP^G = ER^G - EC^G \tag{2}$$

where ER^G =the expected revenue; EC^G =the expected cost; EP^G =the expected profit.

One can derive (2) by substituting the definition of a, b, c, d and e , as well as using the following relationship into (2):

$$E(\min\{Q, D\}) = D - (D - Q)^+, (D - Q)^+ = (D - Q) + (Q - D)^+ \text{ and} \\ (Q - D)^+ = (Q - D) + (D - Q)^+$$

One has

$$ER^G = S\mu - SE(D - Q)^+ + VE(Q - D)^+ \\ - \{A_0 + F + (A_1 + C_p + U)Q + L_s E(D - Q)^+\} \\ EP^G = C_p \{ (a + c - b)\mu - (a + c)E(D - Q)^+ \\ - (c - b - d - e)Q - (b)E(Q - D)^+ \} - A_1 - F \tag{3}$$

In order to maximize (3), the following lemmas from Gallego and Moon [1] are used:

Lemma 1: $E(D - Q)^+ \leq \left(\left[\sigma^2 + (Q - \mu_c)^2 \right]^{\frac{1}{2}} - (Q - \mu_c) \right) / 2$

Lemma 2: $E(Q - D)^+ \leq \left(\left[\sigma^2 + (\mu_c - Q)^2 \right]^{\frac{1}{2}} - (\mu_c - Q) \right) / 2$

The formula of the retailer’s ordering quantity is different from the above derivation and, from (3), can be revised by applying Lemma 1 and Lemma 2:

$$EP^G \geq C_p \left\{ (a + c - b) - (a + c) \frac{\left[\sigma^2 + (Q - \mu_c)^2 \right]^{\frac{1}{2}} - (Q - \mu_c)}{2} \right. \\ \left. - (c - b - d - e)Q - (b) \frac{\left[\sigma^2 + (\mu_c - Q)^2 \right]^{\frac{1}{2}} - (\mu_c - Q)}{2} \right\} - A_1 - F \tag{4}$$

Maximizing the lower bound on (4) is equivalent to minimizing the following function:

$$\begin{aligned} \Theta Ep^G &= \left\{ (c - b - d - e)Q + (a + c) \frac{[\sigma^2 + (Q - \mu_c)^2]^{1/2} - (Q - \mu_c)}{2} \right. \\ &\quad \left. + (b) \frac{[\sigma^2 + (\mu_c - Q)^2]^{1/2} - (\mu_c - Q)}{2} \right\} \\ &= \frac{1}{2} \left\{ (c - b - a - 2(d + e))Q + (a + b + c)[\sigma^2 + (Q - \mu_c)^2]^{1/2} \right. \\ &\quad \left. + (a + c - b)\mu_c \right\} \end{aligned} \tag{5a}$$

Taking the first derivative of ΘEp^G with respect to Q and setting the result to zero, one has

$$\begin{aligned} Q^* &= \frac{1}{1 - \theta} \left\{ \mu_c + \left(\frac{\sigma(R/Z)}{[1 - (R/Z)^2]^{1/2}} \right) \right\}, \\ \text{where } \frac{R}{Z} &= \frac{c - a - b - 2(d + e)}{a + b + c} \end{aligned} \tag{5b}$$

From the relationship between wholesale price and the definition of c , one can realize that the wholesale price influences the ordering decision.

The Manufacturer’s Relevant Costs

The differential equation of the production period for the manufacturer’s inventory level is

$$\frac{d\Psi_s(t_1)}{dt_1} = P - \theta \cdot \Psi_s(t_1) \quad 0 \leq t_1 \leq T_b, \tag{6}$$

Using Spiegel [20] and the various boundary conditions $\Psi_s(0) = 0$ and $\Psi_s(T_p) = Q_v$, the solution of the differential equation is:

$$\Psi_s(t_1) = \frac{P}{\theta} [1 - \exp(-\theta \cdot t_1)]; \quad \Psi_s(T_p) = Q_v = \frac{P}{\theta} [1 - \exp(-\theta \cdot T_p)] \tag{7}$$

Since $\theta \ll 1$ and $T_p \ll 1$, $\exp(-\theta T_p)$ is replaced by $1 - \theta T_p + (\theta T_p)^2 / (2!)$. When $\theta T_p \leq 0.03$, the absolute percentage error is about 0.0464%. It will be smaller for the terms higher than the third term. Therefore, the terms higher than three are neglected. We use the Taylor series to approximate Q_v . One has

$$Q_v = \frac{P}{\theta} \left\{ \theta T_p - (\theta \cdot T_p)^2 / 2 \right\} \tag{8}$$

From (7) and (8), connecting the relationship of the manufacturer and the retailer by setting $Q_v = Q^*$ and $Q^* = Q/(1 - \theta)$, one has the production run time

$$T_p = \frac{1}{\theta} \left(1 - \sqrt{1 - \frac{2\theta Q^*}{P}} \right) = \frac{1}{\theta} \left(1 - \sqrt{1 - \frac{2\theta Q}{P(1 - \theta)}} \right) \tag{9}$$

Property 1

When the deteriorating rate approximate zero, $T_p \rightarrow \frac{Q}{P}$.

Using L'Hospital rule, the proof is completed.

The manufacturer's holding cost considering the deterioration is

$$H \int_0^{T_p} \Psi_S(t_1) dt_1 = H \int_0^{T_p} \frac{P}{\theta} \{1 - \exp(-\theta \cdot t_1)\} = \frac{HP[\theta T_p - 1 + \exp(-\theta T_p)]}{\theta^2} \tag{10a}$$

Since the production is imperfect, we assume an elapsed time until shift is exponentially distributed with a mean of $1/\mu$, one has $f(X) = \mu e^{-\mu X}$ and $1 - F(X) = e^{-\mu X}$, where $1/\mu$ is the arrival rate.

The number of the nonconforming items Z in a production period is given by

$$Z = \begin{cases} \vartheta_1 P T_p & \text{when } X \geq T_p \\ \vartheta_1 P X + \vartheta_2 P(T_p - X), & \text{when } X < T_p \end{cases} \tag{11}$$

Since μ is very small, from (11), the expected number of the nonconforming items is

$$E(Z) = \left\{ \int_0^{T_p} [\vartheta_1 P X + \vartheta_2 P(T_p - X)] f(X) dX + \vartheta_1 P \int_{T_p}^{\infty} f(X) dX \right\} \\ \approx \left[\vartheta_1 P T_p - \frac{(\vartheta_1 - \vartheta_2) P \mu}{2} (T_p)^2 \right] \tag{12}$$

There is a material's delivery-batch effect on finding and reducing the manufacturer's imperfect items because of the frequent JIT deliveries. This phenomenon of JIT benefit is found in practices. As mentioned above, the expected number of non-conforming items' reduction is assumed to be proportion to the number of the material's deliveries, and the rework cost (including the inspection cost) can be derived as

$$RW = C_R \left[E(Z) \left(1 - \sum_{j=1}^M r_j (n_j - 1) \right) \right] \tag{13a}$$

After sale, the product may become unusable. Assuming the hazard rates of the conforming and nonconforming items are of Weibull distribution, the mean failure

rates are $h_1 = \int_0^K v_1(\tau) d\tau = \int_0^K (\lambda_1^{\rho_1} \rho_1 t^{\rho_1 - 1}) dt = (\lambda_1 K)^{\rho_1}$ and

$h_2 = \int_0^K v_2(\tau) d\tau = \int_0^K (\lambda_2^{\rho_2} \rho_2 t^{\rho_2-1}) dt = (\lambda_2 K)^{\rho_2}$, respectively. The free-repair

warranty cost is

$$\begin{aligned}
 PO &= C_w \left\{ \left[E(Z) \left(1 - \sum_{j=1}^M r_j (n_j - 1) \right) \right] h_2 + \left[Q - \left(E(Z) \left[1 - \sum_{j=1}^M r_j (n_j - 1) \right] \right) \right] h_1 \right\} \\
 &= C_w \left\{ \left[\left(\vartheta_1 P T_p - \frac{(\vartheta_1 - \vartheta_2) P \mu}{2} (T_p)^2 \right) \left(1 - \sum_{j=1}^M r_j (n_j - 1) \right) \right] (h_2 - h_1) + P T_p h_1 \right\} \tag{14}
 \end{aligned}$$

The Material’s Relevant Costs of the Manufacturer

The differential equation for the inventory level of the manufacturer’s material j with the boundary condition $\Psi_{mj}(T_p/n_j) = 0$ is

$$\frac{d\Psi_{mj}(t)}{dt} = -\alpha_j P - \theta \Psi_{mj}(t) \quad 0 \leq t \leq T_p/n_j. \tag{15}$$

The solution of the differential equation and for the inventory level of the manufacturer’s material j is:

$$\Psi_{mj}(t_j) = \frac{\alpha_j P}{\theta} \left\{ \exp[\theta(T_p/n_j - t_j)] - 1 \right\} \tag{16}$$

The holding cost of the material j and the material j ’s extra handling cost due to the tax, overhead cost are as follows:

$$\begin{aligned}
 \text{The material } j\text{'s holding cost} &= H_{rj} n_j \int_0^{T_p/n_j} \Psi(t_j) dt_j \\
 &= H_{rj} n_j \int_0^{T_p/n_j} \frac{\alpha_j P}{\theta} \left\{ \exp[\theta(T_p/n_j - t_j)] - 1 \right\} dt_j \tag{17}
 \end{aligned}$$

$$\text{Material } j\text{'s extra handling cost} = \sum_{j=1}^M \left[h_{dmj} \alpha_j n_j (P T_p / n_j) (1 + \theta T_p / (2n_j)) \right] \tag{18}$$

For the manufacturer’s total cost/ profit function under fixed target profit, the problem can be formulated as follows:

Maximize Z= manufacturer’s revenue – (materials’ and manufacturer’s) total cost =

$$\frac{(C_p - G_m)}{1 - \theta} \left\{ \mu_c + \frac{\sigma(R/Z)}{\left[1 - (R/Z)^2 \right]^{1/2}} \right\} - \sum_{j=1}^M \left[h_{dmj} \frac{\alpha_j n_j P}{\theta} \left[\exp\left(\frac{P T_p}{n_j} \right) - 1 \right] \right]$$

$$\begin{aligned}
 & - \left\{ \sum_{j=1}^M \left[n_j C_{mj} + \frac{n_j g_{1j}}{\theta} \cdot \left\{ \frac{-1 - \theta T_p / n_j + \exp(\theta T_p / n_j)}{\theta} \right\} \right. \right. \\
 & + \left. \frac{n_j g_{2j}}{\theta} \left\{ \exp \left[\frac{\theta T_p}{n_j} \right] - 1 \right\} \right. \\
 & \times \left. \left. \left(g_{3j} + \frac{g_{4j} T_p}{2} \right) \right] + C_s + \frac{HP[\theta T_p - 1 + \exp(-\theta T_p)]}{\theta^2} \right. \\
 & + \left. \left[g_5 \left(1 - \sum_{j=1}^M r_j (n_j - 1) \right) + g_7 \right] T_p \right. \\
 & \left. - g_6 \left(1 - \sum_{j=1}^M r_j (n_j - 1) \right) T_p^2 \right\} \tag{19a}
 \end{aligned}$$

Since the optimal ordering decision is determined by retailer, (i.e., the total sale is fixed, and T_p can be derived), we need to determine the minimal total cost of the “manufacturing channel” to maximize the total manufacturing profit. It is equivalent to optimize the problem:

Minimize $TP_G(\underline{n}) =$

$$\begin{aligned}
 & \sum_{j=1}^M \left[h_{dmj} \frac{\alpha_j n_j P}{\theta} \left[\exp \left(\frac{PT_p}{n_j} \right) - 1 \right] \right] \\
 & + \left\{ \sum_{j=1}^M \left[n_j C_{mj} + \frac{n_j g_{1j}}{\theta} \cdot \left\{ \frac{-1 - \theta T_p / n_j + \exp(\theta T_p / n_j)}{\theta} \right\} \right. \right. \\
 & + \left. \frac{n_j g_{2j}}{\theta} \left\{ \exp \left[\frac{\theta T_p}{n_j} \right] - 1 \right\} \right] \times \left(g_{3j} + \frac{g_{4j} T_p}{2} \right) \right] + C_s + \frac{HP[\theta T_p - 1 + \exp(-\theta T_p)]}{\theta^2} \\
 & + \left. \left[g_5 \left(1 - \sum_{j=1}^M r_j (n_j - 1) \right) + g_7 \right] T_p - g_6 \left(1 - \sum_{j=1}^M r_j (n_j - 1) \right) T_p^2 \right\} \tag{19b}
 \end{aligned}$$

where $\underline{n} = (n_1, n_2, \dots, n_M)$, $g_{1j} = \alpha_j PH_{rj}$, $g_{2j} = \alpha_j P$,
 $g_{3j} = L_j (C_{rj} + H_{rj}) + C_{rj}$,
 $g_{4j} = (C_{rj} + H_{rj})$, $g_5 = \{[C_R + C_w (h_2 - h_1)]\vartheta_1 P\}$,
 $g_6 = [C_w (h_2 - h_1) + C_R](\vartheta_1 - \vartheta_2)P\mu/2$, $g_7 = \{C_w Ph_1 + \mu P\}$, $j = (1, 2, \dots, M)$ (19c)

4 Optimization

The purpose of this study is to determine the optimal wholesale price C_p^* for the manufacturer and optimal replenishment number \underline{n}^* for the materials that maximize the total profit. Let $TP_{Gj}(n_j)$ be the cost function for one type of material, one can derive the \underline{n}^* which minimize $TP_G(\underline{n})$.

Proposal 1

If $(C_{mj} + g_6 r_j T_p^2) > g_5 r_j T_p$, there exists n_j^* , $j = 1..M$, that satisfies the condition of

$$n_j^*(n_j^* - 1) \leq \frac{\{[g_{1j} + \theta(h_{dmj} P + g_{2j}(g_{4j} T_p / 4 + g_{3j} / 2))]\}}{2[C_{mj} - r_j T_p (g_5 - g_6 T_p)]} \cdot T_p^2 \leq n_j^*(n_j^* + 1) \tag{20}$$

such that $TP_G(n_1^* - 1, n_2^* - 1, \dots, n_M^* - 1, C_p) \geq TP_G(n_1^*, n_2^*, \dots, n_M^*, C_p)$
 $\leq TP_G(n_1^* + 1, n_2^* + 1, \dots, n_M^* + 1, C_p)$.

Property 2

When the deteriorating rate approximates zero, there is an upper bound of the optimal number of material j 's deliveries when T_p approximates $2C_{mj} / (r_j g_5)$. The optimal condition is

$$n_j(n_j - 1) \leq \frac{\alpha_j PH_{rj}}{\left[\frac{-g_5}{2C_{mj}} \left(r_j - \frac{2g_6 C_{mj}}{g_5} \right)^2 + 2C_{mj} \left(\frac{g_6}{g_5} \right)^2 \right]} \leq n_j(n_j + 1) \tag{21}$$

From Property 2, it is clear that the material's ordering cost, the material's holding cost rate, the production rate, the rework cost and the warranty cost influence the material replenishment policy.

The Sufficient Condition

The sufficient condition for the global optimum is that the Hessian matrix $\nabla^2 TP_G(\underline{n})$ is a positive definite matrix. We can revise the Hessian matrix using the property of $\partial^2 TP_G(\underline{n})/\partial n_i \partial n_j = 0, i \neq j$. Finally, substituting optimal $\underline{n}^* = (n_1^*, n_2^*, \dots, n_M^*)$ into $TP_G(\underline{n})$, one can derive the optimal wholesale price C_p^* under a given target

$$\text{profit: } Q^* \cdot G_m = Q^* \cdot C_p - TP_G(\underline{n}^*) \tag{22}$$

Proof: The revised Hessian Matrix is

$$\nabla^2 TP_G(\underline{n}) = \begin{bmatrix} \frac{\partial^2 TP_G}{\partial n_1^2} & 0 & 0 & \dots & 0 \\ 0 & \frac{\partial^2 TP_G}{\partial n_2^2} & 0 & \dots & 0 \\ 0 & 0 & \ddots & \dots & \vdots \\ \vdots & \vdots & \dots & \frac{\partial^2 TP_G}{\partial n_{M-1}^2} & 0 \\ 0 & 0 & \dots & 0 & \frac{\partial^2 TP_G}{\partial n_M^2} \end{bmatrix}$$

Since $\partial^2 TP_G/\partial n_j^2 > 0$ and matrix $[]_{j \times j}$ is positive, the $\nabla^2 TP_G(\underline{n})$ is a positive definite matrix.

Solution Procedure:

Due to the complexity of solving the algebraic solution with respect to the total profit function, this paper proposes a solution procedure to derive the optimal values for the proposed model.

- Step 1: Input the parameters of the problem.
- Step 2: If the problem satisfies Proposal 1 for all types of materials, go to Step 3; otherwise, go to step 5.
- Step 3: Use (20) to solve for the optimal value of n_j^* . If all the n_j^* satisfy the sufficient condition of the optimal solution, the solution of $\underline{n}^* = (n_1^*, n_2^*, \dots, n_M^*)$ is the optimal solution; Otherwise, go to step 5.
- Step 4: Use (5b) to compute production lot size. Then derive the total cost using (19a) and compute C_p^*
- Step 5: Stop.

5 Numerical Example

The theory can be illustrated by the numerical example (using the inputs in Table 3). The optimal solution for the multiple materials case is presented in the Table 4. The optimal solution considering varying deteriorating rate is shown in Table 5.

Table 3. Inputs for example

μ_c	800	C_R	40	μ	0.015	h_{dm2}	13
σ	70	C_w	90	K	2	C_{r1}	4
a	0.64	G_m	25	λ_1	0.01	C_{r2}	4.6
b	0.6	u	2	λ_2	0.012	H_{r1}	3
c	0.58	H	4.5	ρ_1	0.8	H_{r2}	4
d	0.04	L_1	0.0027	ρ_2	0.8	α_1	2
e	0.05	L_2	0.0025	A_0	230	α_2	3
C_S	1000	C_{m1}	300	F	600	r_1	0.001
P	1200	C_{m2}	310	θ_1	1/360	r_2	0.001
θ	0.01	h_{dm1}	10	θ_2	1/240		

Table 4. The optimal solution of the example

Optimal solution	$(n_1^*, n_2^*) = (2.26, 3.14) = (2, 3)$	Ordering quantities = 771.29
	$C_p^* = 76.62$	Total cost $TP_G(\underline{n}^*) = 39671.64$

Table 5. The optimal solution for varying deteriorating rate

θ	(n_1^*, n_2^*)	C_p^*	Ordering quantities	Total cost $TP_G(\underline{n}^*)$
0.10	(3,4)	79.77	848.42	44527.24
0.01	(2,3)	76.62	771.58	39671.64
0.001	(2,3)	76.35	764.35	39233.91
0.0001	(2,3)	76.32	763.66	39190.68
0.00001	(2,3)	76.32	763.59	39186.36

6 Conclusion

This paper proposes a two-echelon distribution-free integrated production-inventory deteriorating model with imperfect process. The necessary and sufficient conditions for the optimal solution are derived and a solution procedure is developed to give the management insight in determining the optimal wholesale price and the optimal replenishing cycle. In this study, we derive an upper bound for the optimal number of material's JIT deliveries.

References

1. Gallego, G., Moon, I.: The distribution-free newsboy problem—review and extensions. J. Oper. Res. Soc. 44(6) (1993) 734–825.
2. Lau, H.S., Lau, A.H.L.: Manufacturer's pricing strategy and return policy for a single-period commodity. Eur. J. Oper. Res. 116 (1999) 291–304.

3. Lau, A.H.L., Lau, H.S.: The effects of reducing demand uncertainty in a manufacturer-retailer channel for single-period products. *Comput. Oper. Res.* 29 (2002) 1583–1602.
4. Lariviere, M.A., Porteus, E.L.: Selling to the newsvendor: an analysis of price-only contracts. *Mfg. & Service Oper. Management* 3(4) (2001) 293–305.
5. Shang, K.H., Song, J.S.: Newsvendor bounds and heuristic for optimal policies in serial supply chains. *Management Sci.* 49(5) (2003) 618–638.
6. Moon, I., Choi, S.: Distribution free newsboy problem with balking. *J. Oper. Res. Soc.* 46(4) (1995) 537–542.
7. Moon, I., Silver, E.A.: The multi-item newsvendor problem with a budget constraint and fixed ordering costs. *J. Oper. Res. Soc.* 51 (5) (2000) 602–608.
8. Ghare, P.M., Schrader, S.F.: A model for exponentially decaying inventory. *J. Ind. Eng.* 14(5) (1963) 238-243
9. Covert, R.P., Philip, G.C.: An EOQ model for items with Weibull distribution deterioration. *AIIE Trans* 5 (1973) 323-326.
10. Raafat, F., Wolfe, P.M., Eldin, H.K.: An inventory model for deteriorating items. *Comput. Ind. Eng.* 20 (1991) 89-94
11. Wee, H.M.: Economic production lot size model for deteriorating items with partial back-ordering. *Comput. Ind. Eng.* 24(3) (1993) 449-458.
12. Wee, H.M., Jong, J.F.: An integrated multi-lot-size production inventory model for deteriorating items. *Management and System* 5(1) (1998) 97-114.
13. Yang, P.C., Wee, H.M.: An integrated multi-lot-size production inventory model for deteriorating item. *Comput. Oper. Res.* 30(5) (2003) 671-682.
14. Nassimbeni, G.: Factors underlying operational purchasing practices: Results of a research. *Int. J. Prod. Econ.* 42 (1995) 275-288.
15. Reese, J., Geisel, R.: A comparison of current practice in German manufacturing industries. *E. J. Purchasing & supply management* 42 (1997) 275-288.
16. Lee, H.L., Rosenblatt, M.J.: Simultaneous determination of production cycle and inspection schedules in a production system. *Management Sci.* 33 (1987) 1125-1136.
17. Wang, C.H., Sheu, S.H.: Optimal lot size for products under free-repair warranty. *Eur. J. Oper. Res.* 149 (2003)131-141.
18. Wang, C.H.: The impact of a free-repair warranty policy on EMQ model for imperfect production system. *Comput Oper. Res.* 31 (2004) 2021-2035.
19. Yeh, R.H., Ho, W.T., Teng, S.T.: Optimal production run length for products sold with warranty. *Eur. J. Oper. Res.* 120 (2005) 575-582.
20. Spiegel, M.R.: *Applied differential equations*. Prentice-Hall, Englewood Cliffs. N.J (1960).

Mathematical Modeling and Tabu Search Heuristic for the Traveling Tournament Problem

Jin Ho Lee, Young Hoon Lee, and Yun Ho Lee

Department of Information and Industrial Engineering, Yonsei University,
134 Shinchon-Dong, Seodaemun-Gu, Seoul 120-749, Korea
{younggh, jinho7956, yuno80}@yonsei.ac.kr

Abstract. As professional sports have become big businesses all over the world, many researches with respect to sports scheduling problem have been worked over the last two decades. The traveling tournament problem (TTP) is defined as minimizing total traveling distance for all teams in the league. In this study, a mathematical model for the TTP is presented. This model is formulated using an integer programming (IP). In order to solve practical problems with large size of teams, a tabu search heuristic is suggested. Also, the concepts of alternation and intimacy were introduced for effective neighborhood search. Experiments with several instances are tested to evaluate their performances. It was shown that the proposed heuristic shows good performances with computational efficiency.

1 Introduction

Recently, as sports industries have become big businesses all around the world, high income has been obtained through many sports leagues. One key to such high income levels is the schedule the teams play. Thus, many researches concerning sports scheduling problem have been worked in Operations Research over the two decades.

The term “Traveling Tournament Problem” (TTP) with respect to sports scheduling problem is defined as minimizing total traveling distance for all teams in the league. The TTP represents the fundamental issues involved in creating a schedule for sports leagues where the amount of team travel are an issue. For many of these leagues, the scheduling problem includes a myriad of constraints based on thousands of games and hundreds of team idiosyncrasies that vary in their content and importance from year to year, so instances of this problem seem to be very difficult to solve even for very small cases. Thus, the TTP is considered as an interesting challenge for combinatorial optimization techniques. In this paper, a mathematical model for the TTP is presented and a tabu search heuristic method is proposed in order to solve the instances where there are a large number of teams involved.

2 Literature Review

The sports scheduling research literature has focused on league scheduling over the past two decades. James and John [8] presented a model that reduces travel cost and

player fatigue using an integer programming approach to the National Basketball Association (NBA). Fleurent and Ferland [6] allocated the games using an integer programming to the National Hockey League (NHL), and Costa [3] used a tabu search algorithm for the NHL. Schreuder [12] constructed a timetable for the Dutch professional soccer leagues that minimized the number of schedule breaks. Russell and Leung [11] developed a schedule for the Texas Baseball League that satisfied stadium availability and various timing restrictions with the objective of minimizing travel costs. Nemhauser and Trick [10] created a schedule for the Atlantic Coast Conference men's basketball season, taking into consideration numerous conflicting requirements and preferences.

Easton *et al.* [4] defined the traveling tournament problem as having the objective of minimizing total traveling distances for all teams in the league. Additionally, Easton *et al.* [5] presented an optimal solution for an instance of four, six and eight teams through the Branch-and-Price algorithm using integer programming and constraint programming. Benoist *et al.* [2] presented a study of hybrid algorithms combining lagrangian relaxation and constraint programming on a round-robin assignment and travel optimization. Anagnostopoulos *et al.* [1] achieved good results with a simulated annealing algorithm that explores both feasible and infeasible schedules for the TTP, searching large neighborhoods with complex moves. Lim *et al.* [9] improved the quality of solutions for instances of large size using a simulated annealing and hill-climbing algorithm for the TTP.

As described above, a lot of studies have been conducted in this field. However, providing the optimal schedule for instances of more than eight teams still remains an open problem.

This paper is organized as follows: In Chapter 3, the traveling tournament problem is described, and then a mathematical model is presented. In Chapter 4, a heuristic method using tabu search is proposed. In Chapter 5, the proposed heuristic is evaluated on the basis of computational results. Finally, concluding remarks and recommendations for future researches are addressed in chapter 6.

3 Problem Description and Mathematical Modeling

This chapter describes the traveling tournament problem, which has the objective of minimizing total distance traveled by all teams in the league. The chapter also deals with several of the complex constraints, which consist of double round-robin tournament constraints, consecutive constraints and no-repeat constraints representatively. And then a mathematical model using an integer programming is presented.

3.1 Definition of Traveling Tournament Problem

Easton *et al* [4] introduced and defined the traveling tournament problem (TTP), and the definition is as follows: "Given n teams and n evens, a round-robin tournament is a tournament among teams so that every team plays every other team. Such a tournament has $n - 1$ slots during which $n / 2$ games are played. For each game, one team is denoted the home team and its opponent is the away team. As suggested by the name, the game is held at the venue of the home team. A double round-robin tournament has

$2(n - 1)$ slots and has every pair of teams played twice, once at home and once away for each teams.”

For the problem, distances between team sites are given by n by n matrix. Assuming equality of distances to and from sites, this matrix is symmetric. Each team begins the tournament at its home site to which it must return at the end of the tournament. Also, when a team plays an away games, a team travels from one away venue to the next directly. The cost to each team is the total distance traveled starting from its home site and ending back there on completion of its scheduled games. In summery, the input data and output of the TTP is as follows:

Input: n , the number of teams, D , an n by n integer symmetric distance matrix.

Output: A double round-robin tournament on the n teams such that

1. All constraints on the tournament are satisfied, and
2. The total distance traveled by the teams is minimized.

The TTP consists of a large number of constraints, so it has been recognized as a difficult problem to solve even for very small cases, but there are two basic requirements. The first is a feasibility issue that the home and away pattern must be sufficiently varied so as to avoid long home stands and road trips. A road trip is defined as a series of consecutive away games. Similarly, a home stand is the number of home games in the series. The second is an optimality that minimizes total traveling distance for all teams. The key to the TTP is a conflict between minimizing travel distances and feasibility constraints on the home/away pattern. A solution to the TTP must satisfy the following several constraints:

1. Double round-robin constraints

Each pair of teams, for example a pair of teams, A and B, play exactly twice-once at A’s home site and once at B’s home site. Thus, there is the exit total $2(n - 1)$ rounds and $n / 2$ games are played in each round.

2. Consecutive constraints

For each team, no more than three consecutive home or three consecutive away games are allowed. In other words, more than three consecutive home stands or road trips are forbidden by these constraints.

3. No-repeat constraints

For any pair of teams, for example a pair of teams, A and B, after A and B play at A’s home site, A and B cannot immediately play at B’s home site in next round.

3.2 Mathematical Modeling for the TTP

In this paper, a mathematical model for the TTP is presented. This model was formulated using an integer programming (IP). Also, this is an extended model from James and John’s one [8] and the details are as follows:

Notation

i : the index for the team’s site or venue ($i = 1, 2, \dots, n$)

j : the index for the team’s site or venue ($j = 1, 2, \dots, n$)

k : the index for the team’s site or venue ($k = 1, 2, \dots, n$)

t : the index for the round ($t = 1, 2, \dots, 2n-2$)

d_{ij} : distance between team i ’s site and team j ’s site

Decision Variable

$Y_{i,j,k,t}$: 1 if team i moves from team j 's site to team k 's site at round t , 0 otherwise

<IP Formulation>

$$\text{Minimize } \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{t=1}^{2n-2} d_{jk} \times Y_{i,j,k,t} \tag{1}$$

subject to

$$\sum_{j=1}^n \sum_{t=1}^{2n-2} Y_{i,j,i,t} = n-1 \quad \forall i \tag{2}$$

$$\sum_{j=1}^n \sum_{t=1}^{2n-2} Y_{i,j,k,t} = 1 \quad \forall i, k (k \neq i) \tag{3}$$

$$\sum_{i=1}^n \sum_{j=1}^n Y_{i,j,k,t} \leq 2 \quad \forall i, j, t (i \neq j) \tag{4}$$

$$\sum_{j=1}^n Y_{i,j,k,t} \leq \sum_{j=1}^n Y_{k,j,k,t} \quad \forall i, k, t \tag{5}$$

$$\sum_{j=1}^n Y_{i,j,k,t} = \sum_{j=1}^n Y_{i,k,j,t+1} \quad \forall i, k, t \tag{6}$$

$$\sum_{j=1}^n Y_{i,j,i,1} = 1 \quad \forall i \tag{7}$$

$$\sum_{j=1}^n Y_{i,j,i,2n-1} = 1 \quad \forall i \tag{8}$$

$$Y_{i,k,i,t} + Y_{k,k,i,t} \leq 1 \quad \forall i, k, t (i \neq k) \tag{9}$$

$$\sum_t^{t+3} Y_{i,i,i,t} \leq 3 \quad \forall i, t = 1, 2, \dots, 2n-5 \tag{10}$$

$$\sum_{\substack{j=1 \\ \neq i}}^n \sum_{\substack{k=1 \\ \neq i}}^n \sum_t^{t+3} Y_{i,j,k,t} \leq 3 \quad \forall i, t = 1, 2, \dots, 2n-5 \tag{11}$$

The objective function (1) is to minimize total traveling distance for all teams in the league. Equation (2) is the constraint for the number of $n - 1$ games must be played in its home site, and equation (3) means that each team must play once with any other teams in each opponent's home site. Equation (4) represents that no more than two teams can locate in one venue simultaneously, and equation (5) means that if an away team visits, a home team must be located in its own home site. Thus, two teams can be located in one venue simultaneously or no team is located according to constraints (4) and (5). Equation (6) is moving sequence constraints that if team i was in team j 's home at round t , team i must start in team j 's home at round $t+1$. Equation (7) and (8) are the constraints that all teams must start in their home sites at the beginning of the league and return to their homes at the end of the league. Equation (9) is the no-repeat constraints, and finally equation (10) and (11) are the consecutive constraints.

A mathematical model is tested using an ILOG OPL Studio 3.6.1. However, this does not show optimal solutions for the instances of more than four teams because of a large number of constraints and variables. For an instance of just 4 teams, an optimal solution is solved, and the optimal value for instances of the NL4 and CIRC4 are 8276 and 20. Thus, in the next chapter a tabu search heuristic is proposed in order to solve the practical problems with large size of teams.

4 Tabu Search Heuristic

The Tabu Search (TS) algorithm is an iterative improvement approach designed to avoid terminating prematurely at a local optimum for combinatorial optimization problems. In the TTP, it is difficult to maintain the feasibility of solutions for the consecutive and no-repeat constraints. Thus, it is important to generate good neighborhood, and to search unvisited neighborhood. One issue for tabu search is how to represent solution and how to generate neighborhood. Another is to keep the characteristic of good solution. In this point of view, our algorithm was designed.

4.1 Representation of Solutions and Neighborhoods

In this paper, a schedule is represented by a time-table indicating the opponents of the teams. Each row corresponds to a team and each column corresponds to a round. The opponent of team i at round t is given by the absolute value of element (i, t) . If (i, t) is positive, the game takes place at i 's home, otherwise at i 's opponent home. This representation of schedules is designed by Anagnostopoulos *et al.* [9]. The example of the schedule for 6 teams can be shown in Table 1.

Table 1. Example of the schedule for 6 teams

Team / Round	1	2	3	4	5	6	7	8	9	10
1	3	-6	4	2	-4	-5	5	-3	6	-2
2	-4	4	-5	-1	6	-6	3	5	-3	1
3	-1	-5	-6	6	5	4	-2	1	2	-4
4	2	-2	-1	5	1	-3	6	-6	-5	3
5	6	3	2	-4	-3	1	-1	-2	4	-6
6	-5	1	3	-3	-2	2	-4	4	-1	5

In table 1, while team 1 has home games with team 3, 4, 2, 5 and 6 at round 1, 3, 4, 7 and 9, team 1 has away games with team 6, 4, 5, 3 and 2 in round 2, 5, 6, 8, and 10. Thus, total distance for team 1 is as follows:

$$\text{Total distance of team 1} = d_{16} + d_{61} + d_{14} + d_{45} + d_{51} + d_{13} + d_{31} + d_{12} + d_{21}$$

The neighborhood generations can be obtained by applying five types of moves proposed by Anagnostopoulos *et al.* [1].

4.2 Solution Strategy for the TTP

In order to keep the good solution’s characteristics and search the effective neighborhoods, in this paper the terms “*alternation*” and “*intimacy*” are defined. The details are as follows:

Alternation: *Alternation* means the number of home-away or away-home conversions to one team. The elite schedule has small alternations.

Intimacy: Between the two adjacent rounds, the *intimacy* represents the number of home-home patterns and away-away patterns. The intimacy is use in order to evaluate the each pair of column(round).

In a specific team’s schedule, if the team has high numbers of alternations, it should move lot. Additionally, between the two specific adjacent rounds, low intimacy needs to be improved through the neighborhood search. Alternation and intimacy is shown in Table 2.

Table 2. Alternation and intimacy

T/R	1	2	3	4	5	6	7	8	9	10
						•				
						•				
<i>i</i>	-2	-4	-1	6	5	4	-6	1	2	-5
						•				
						•				

T/R			<i>t</i>	<i>t</i> +1		
1			-6	4		
2			4	-5		
3	•	•	-5	-6		•
4			-2	-1		
5			3	2		
6			1	3		

In the left of Table 2, team *i* has 4 alternations {(-1, 6), (4, -6), (-6, 1), (2, -5)}, and in the right of Table 2, intimacy between round *t* and *t*+1 is 4 {(-5, -6), (-2, -1), (3, 2), (1, 3)}. In the TTP, generating all the neighbors of the current solution needs too much time. Thus, alternation and intimacy is used to search good neighbors effectively. For each team, alternation is evaluated and then 3 teams, which have the maximum alternations, are selected. Also, for each adjacent pair of rounds, intimacy is evaluated and then until remaining 3 or 4 rounds, minimum pair of rounds is deleted in the candidates.

4.3 Tabu Search Algorithm for the TTP

The basic search procedure is driven by a tabu search meta-heuristic [7]. The initial solution is generated randomly, using constraint programming. After evaluating the

alternation and intimacy in the current solution, 3 max-alternation teams and 3 or 4 min-intimacy rounds are selected. Then the neighbors are generated by applying *SwapHome* and *SwapTeam* using 3 max-alternation teams and *SwapRound* using 3 or 4 min-intimacy. Finally, *PartialSwapRound* and *PartialSwapTeam* find the neighbors, using both 3 max-alternation teams and 3 or 4 min-intimacy rounds. The outline of neighborhood generation is shown in Figure 1.

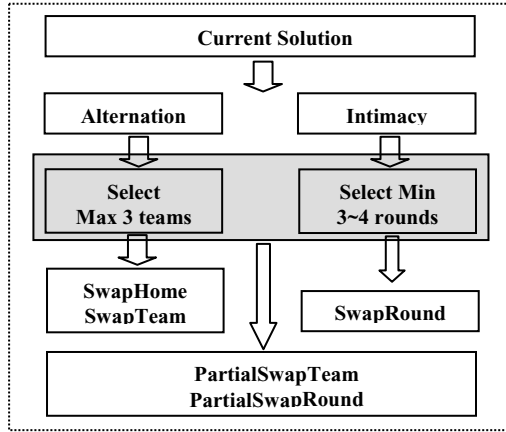


Fig. 1. The outline of neighborhood generation

If a neighborhood is infeasible, it is not considered. The stop criterion is based on the number of iterations. However, when it could not find any improvement for the current best solution, it is evaluated from second best solution to fifth best solution in the same way. If the evaluation of fifth best solution is not improved, then the procedure is finally stopped. The parameters for these experiments are set as 10,000 max iterations, 5 tabu list size.

The steps of our proposed tabu search are progressed by general tabu search meta-heuristic approach. However in the point of neighborhood search, this algorithm finds the good solution’s characteristic through alternation and intimacy. Also, since selected teams and rounds are only candidates with respect to neighborhood, much time is not spent in neighborhood generation.

5 Computational Experiments

The proposed algorithm is tested on two different sets of instances. Namely, experiments with the the real distance NLx instances and the circular instances CIRCx are tested. The detail of these two sets is shown in [13]. In each instance, 10 runs of the algorithm are implemented, recording the best found solution and the total running time of the procedure. The experiments were performed on a Pentium 4, CPU 2.8GHz, 512MB RAM and the algorithm was coded in C language.

The results and comparison of the proposed tabu search heuristic with the best solutions of other previous researches are shown in Table 3, and the comparisons of computational time with previous meta-heuristic approaches are in Table 4.

Table 3. Results and comparison of the best solution

Instance	1)	2)	3)	4)	Our Results
NL6	23916	N.A.	23916	N.A.	23916
NL8	42517	39721	39721	N.A.	39721
NL10	68691	59583	59821	59436	59583
NL12	143655	111248	115089	112298	111483
NL14	301113	189766	196363	190056	190174
NL16	437273	267194	274673	272902	276520
CIRC6	N.A.	N.A.	64	N.A.	64
CIRC8	N.A.	N.A.	132	N.A.	132
CIRC10	N.A.	N.A.	246	242	244
CIRC12	N.A.	N.A.	408	N.A.	416
CIRC14	N.A.	N.A.	654	N.A.	662
CIRC16	N.A.	N.A.	928	N.A.	976
CIRC18	N.A.	N.A.	1356	N.A.	1364
CIRC20	N.A.	N.A.	1842	N.A.	1896

- 1) Lagrange relaxation and constraint programming by Benoit Rottemberg *et al.*, [2]
 - 2) Simulated annealing by Anagnostopoulos *et al.*, [1]
 - 3) Simulated annealing and hill-climbing by Lim *et al.*, [9]
 - 4) By Glenn Langford, [13]
- ※ Bold type indicates the best solution so far, and N.A. indicates “not applicable”.

Table 4. Comparison of the computational time

Instance	Simulated annealing by [1]		Simulated annealing & Hill-climbing by [9]		Proposed Tabu search	
	Min	Mean	Min	Mean	Min	Mean
NL6	N.A.	N.A.	577	821.2	308	379.1
NL8	596.6	1639.3	3224	4106.8	1411	2204.5
NL10	8084.2	40268.6	8557	40289.4	3427	6135
NL12	28526	68505.3	10355	54026.5	7329	26751.2
NL14	418358.2	233578.4	21978	59019.7	15674	41600.1
NL16	344633.4	192086.6	50767	83287.0	48294	71088.3
CIRC6	N.A.	N.A.	594	648.3	412	464
CIRC8	N.A.	N.A.	3340	3729.9	1549	2021.3
CIRC10	N.A.	N.A.	5886	23022.8	3996	8476
CIRC12	N.A.	N.A.	8709	37947.1	7034	28457.8
CIRC14	N.A.	N.A.	11022	53751.0	10426	36628.8
CIRC16	N.A.	N.A.	13261	56036.9	12467	44552.2
CIRC18	N.A.	N.A.	16199	63872.2	14234	55074.2
CIRC20	N.A.	N.A.	50303	68807.0	24715	60815.4

It can be observed in Table 3 and 4 that our performances with respect to objective value are not better than previous best. However, the suggested tabu search shows the good performance regarding computation time. The computation time of the suggested algorithm is shorter than that of [1] and [9]. And, partially the better solutions than [1] and [9] are found.

As shown in the result of experiments, the selection of candidates by application of alternation and intimacy makes this algorithm powerful in the point of performance with computational efficiency for problems.

6 Conclusion

Sports league scheduling has received considerable attention in recent years, since these applications involve significant revenues for television networks and generate challenging combinatorial optimization problems. This paper deals with the traveling tournament problem (TTP) proposed in [4].

Although several recent researches with respect to the TTP propose the heuristic approaches which involve meta-heuristic and hybrid algorithm, the general mathematical model to the TTP is not presented. Thus, this paper presents a mathematical model using integer programming. Additionally, we suggest a tabu search heuristic algorithm, and introduce the mechanism which represents the characteristics of elite solutions. Through the experiments, the proposed heuristic proved that it shows the good performances with computational efficiency. Thus, our mechanism to select candidates to neighborhood generation showed that it makes the suggested heuristic effective in the point of computational time.

In further research, it would be interesting to use the advantages of each meta-heuristic in the hybrid approach. And, if a nice neighbor search technique is presented, it may also improve the efficiency of the algorithm.

References

1. Anagnostopoulos, A., Michael, L., V. Hentenryck, P. and Vergados, Y.: A simulated annealing approach to the traveling tournament problem. In Proceeding of CP-AI-OR (2003)
2. Benoist, T., Laburthe, F. and Rottembourg, B.: Lagrangian relaxation and constraint programming collaborative schemes for traveling tournament problem. CP-AI-OR, Wye College (2001) 15-26
3. Costa, D.: An evolutionary tabu search algorithm and the NHL scheduling problem. *INFOR*, Vol. 3(33). (1995) 161-178
4. Easton, K., Nemhauser, G., and Trick, M.: The traveling tournament problem description and benchmarks. In Proceeding of the 7th International Conference on the Principle and Practice of Constraint Programming, Paphos, Cyprus (2001) 580-589
5. Easton, K., Nemhauser, G., and Trick, M.: Solving the traveling tournament problem: a combined integer programming and constraint programming approach. *Lecture Note in Computer Science*, Vol. 2740. (2003) 100-109
6. Ferland, J. A. and Fleurent, C.: Computer aided scheduling for a sport league. *INFOR*, Vol. 29(1). 14-25
7. Glover, F. and Laguna, M.: *Tabu search*. Kluwer Academic Publishers, (1997)
8. James, C. B. and John, R. B.: Reducing traveling costs and player fatigue in National Basketball Association. *The Institute of Management Science*, Vol. 10(3). (1980) 98-102

9. Lim, A., Rodrigues, B. and Zhang, X.: A simulated annealing and hill-climbing algorithm for the traveling tournament problem. *European Journal of Operational Research*, Vol. 131. (2005) 78-94
10. Nemhauser, G. L. and Trick, M. A.: Scheduling a major college basketball conference. *Operations Research*, Vol. 46(1). (1998) 1-8
11. Russel, R. A. and Leung, J. M. Y.: Devising a cost effective schedule for a baseball league. *Operations Research*, Vol. 42(4). (1994) 614-625
12. Schreuder, J. A. M.: Combinatorial aspects of construction of competition Dutch professional football leagues. *Discrete Applied Mathematics*, Vol. 35(3). (1992) 301-312
13. Trick, M.: Challenge traveling tournament problem. <http://mat.gsia.cmu.edu/TTP/>, (2004)

An Integrated Production-Inventory Model for Deteriorating Items with Imperfect Quality and Shortage Backordering Considerations

H.M. Wee¹, Jonas C.P. Yu², and K.J. Wang³

¹ Department of Industrial Engineering, Chung Yuan Christian University,
Chungli 32023, Taiwan
weehm@cycu.edu.tw

² Logistics Management Department, Takming College, Taipei 114, Taiwan

³ Department of Business Administration, National Dong Hwa University, Hualien, Taiwan

Abstract. In this study we present a production-inventory model for deteriorating item with vendor-buyer integration. A periodic delivery policy for a vendor and a production-inventory model with imperfect quality for a buyer are established. Such implicit assumptions (deteriorating items, imperfect quality) are reasonable in view of the fact that poor-quality items do exist during production. Defective items are picked up during the screening process. Shortages are completely backordered. The study shows that our model is a generalization of the models in current literatures. An algorithm and numerical analysis are given to illustrate the proposed solution procedure. Computational results indicate that our model leads to a more realistic result.

1 Introduction

Since the development of the economic order quantity (EOQ) more than four decades ago, a substantial amount of researches have been conducted in the area of inventory lot sizing. However, one of the weaknesses of most researches is the unrealistic assumption of perfect quality items [25]. Cheng [2] proposed an EOQ model with demand-dependent unit production cost and imperfect production process. He proposed a general power function to model the relationship between unit production cost, demand rate and process reliability. Cheng formulated this inventory decision problem as a geometric problem (GP), and applied the theories of GP to derive a closed-form optimal solution. Zhang and Gerchak [30] considered a joint lot sizing and inspection policy, for an EOQ model with a random proportion of defective units. They considered a model where the defective units are replaced by non-defective ones. Rosenblatt and Lee [24] considered the presence of defective products in a small lot size replenishment policy. They assumed that the defective rate from the beginning in-control state until the process goes out of control increased exponentially. The defective items can be reworked instantaneously and kept in stock. Rosenblatt and Lee concluded that the presence of defective products resulted in smaller lot sizes. Schwaller [26] presented a procedure to extend EOQ models by assuming that the defectives of a known proportion were present in the incoming lots,

and that fixed and variable inspection costs were incurred in finding and removing the items. Porteus [22] incorporated the effect of defective items in the basic EOQ model and invested in process quality improvement. He assumed a probability q would go out of control during production.

Salameh and Jaber [25] presented a modified inventory model for imperfect quality items. They considered poor-quality items are sold as a single batch by the end of the 100% screening process. Rosenblatt and Lee [24] showed that reducing the lot size quantity increased the average percentage of imperfect quality items. The reasonable explanation is that Rosenblatt and Lee [24] assumed defective items were reworked instantaneously and kept in stock. This increases the holding cost that results in lower lot sizes, whereas in this paper, imperfect quality items are withdrawn from stock resulting in lower holding cost and larger lot sizes. Goyal and Gardenas-Barron [10] extended Salameh and Jaber's model and presented a practical approach to determine EPQ for items with imperfect quality. The approach suggested in their study results in nearly a zero penalty as compared to Salameh and Jaber. Later, Goyal *et al.* [11] extended the model of Goyal and Gardenas-Barron [10] to consider vendor-buyer integration. Chung and Hou [3] developed a model to determine an optimal run time for a deteriorating production system with shortages. They assumed the elapsed time is random between the production process shifts.

Recently, Wee and Yu [28] extended the approach by Salameh and Jaber and considered permissible shortage backordering. They found that the traditional EOQ and Salameh and Jaber's modified EPQ/EOQ model are both special cases of the proposed model when the backordering cost is very large. In this paper, the influence of imperfect quality and deterioration is taken into account. Imperfect quality is the result of imperfect machines and processes. Deterioration occurs because many agricultural products, gasoline and medicine do not have constant utility during storage. The distribution of time to deterioration of the item follows the exponential distribution.

Ongoing deteriorating inventory has been studied by several authors in recent decades. Ghare and Schrader [12] were the first authors to consider ongoing deterioration of inventory. They have developed an EOQ model for items with an exponentially decaying inventory. Elsayed and Terasi [5] proposed a deteriorating production-inventory model with Weibull distribution and permissible shortage. Kang and Kim [20] proposed an exponentially deteriorating model considering the price and production level. An exponentially deteriorating production-inventory model with permissible shortage is presented in [23]. Other authors such as Dave [4] and Heng *et al.* [16] assumed either instantaneous or finite production rate with different assumptions on the patterns of deterioration. Yang and Wee [29] developed an integrated economic ordering policy of deteriorating items for a vendor and multiple-buyers.

Collaboration of enterprises, especially in terms of developing strategies, is vital in reducing the overall cost of the enterprise. This is because decision made independently by one player will not result in global optimum. Global optimization will only be realized if the perspectives of all players are considered. One of the advantages of applying joint economic lot size models (JELS) is being able to generate lower total inventory relevant cost for the system so that the net benefit can be shared by both parties. The JELS approach has been studied for years. Goyal [6]

was the first to introduce an integrated inventory policy for the single-supplier single-customer problem. He showed that his integrated policy results in minimum joint variable cost for the supplier and the customer. Banerjee [1] developed a joint economic lot size model with lot-for-lot policy for a single-buyer single-vendor system by combining two EOQ models from the buyer and vendor. In his model, he assumed that the vendor makes the production setup as long as the buyer places an order and supplies on a lot for lot basis. He also showed that his JELS model has minimum joint total relevant cost by considering both the buyer and the vendor at the same time. Later Goyal [7] generalized Banerjee's model by relaxing the assumption of the vendor's lot-for-lot policy. He pointed out that the vendor could possibly produce a lot that can supply an integer number of orders from the buyer. Nevertheless the model restricts shipments cannot be triggered before the whole production batch is completed. A review of previous models on buyer-vendor integration until 1990 refers to [8].

Lu [21] developed an algorithm which derived an optimal solution to the single-vendor single-buyer problem, when the delivery quantity to the buyer at each replenishment was identical. Lu's model is synchronous, allowing shipment to take place during production. The model proposed by Halm and Yano [14] is also synchronous, aiming to minimize the manufacturer's and buyer's inventory holding cost, manufacturer's setup cost, as well as the transportation cost. Halm and Yano advocates that for the single-buyer single-item problem, the optimal solution has the property that the production interval is an integer multiple of the delivery interval. Halm and Yano [15] further extend the model to the single-machine multi-component problem. A heuristic procedure was therefore developed to find both the "just-in-time" production runs and the delivery schedule. Goyal [9] relaxes the identical shipment constraint, allowing the quantity of successive shipments to be different in an increasing fashion by a fixed factor of production rate to demand rate. The policy is to deliver whatever that is produced at the replenished time by using the same example as Lu [21]. Goyal shows that a different shipment policy could result in a better solution. Ha and Kim [13] analyze the integration between the buyer and the supplier, and developed a mathematical model using the geometric method.

Though both Goyal's [9] and Hill's [17] models illustrate that delivery in "what is produced" policy is better than delivery in "identical shipment" policy, Viswanathan [27] shows that neither strategy obtains the best solution for all possible problem parameters. More recently, Hill [18] derived a globally optimal batching and shipping policy for the single-vendor single-buyer integrated production-inventory problem. Hoque and Goyal [19] proposed an optimal policy for a single-vendor single-buyer integrated production-inventory system with a limited capacity of transport equipment.

In this study, product deterioration and vendor-buyer integration are considered simultaneously. We propose a production-inventory model for an on-going deterioration item with partial backordering and imperfect quality. Shortages due to imperfect items are completely backordered. This is because not all customers are willing to wait for a new replenishment of stock. Customers encountering shortages will respond differently according to the type of commodities and market environment. In real world, complete backordering is likely only in a monopolistic market. An illustrative example and sensitivity analysis are given to validate the inventory model.

2 Notation and Assumptions

The following notation is used:

$I(t) =$	Inventory level at time t , $0 \leq t \leq T$;
$I_m =$	Maximum inventory level;
$I_s =$	Inventory level at the end of time t_l ;
$T_{v1} =$	The time length of the production stage;
$T_{v2} =$	The time length of the non-production stage;
$R =$	The total production quantity;
$Q =$	Order size;
$D =$	Demand rate for buyer;
$X =$	Screening rate for buyer;
$c =$	Purchasing cost per unit for buyer;
$h =$	Carrying cost per unit per unit time for buyer;
$d =$	Deterioration cost per unit for buyer;
$K =$	Ordering cost per order for buyer;
$\theta =$	Deterioration rate;
$C_{vh} =$	Carrying cost per unit per unit time for vendor;
$C_{vd} =$	Deterioration cost per unit for vendor;
$p =$	Defective percentage in demands for vendor;
$f(p) =$	Probability density function of p ;
$x =$	Screening cost per unit for buyer;
$I_b =$	Total shortage demand (units/cycle);
$b =$	Backordering cost per unit for buyer;
$* =$	Superscript representing optimal value;

The mathematical models presented in this study have the following assumptions:

- (1) A single item with constant deteriorating rate of the on-hand inventory is considered.
- (2) Demand rate is a continuous known constant.
- (3) Lead-time is a known constant.
- (4) Defective items are independent of deterioration.
- (5) Replenishment is instantaneous.
- (6) Screening process and demand proceeds simultaneously.
- (7) Defective percentage, p , has a uniform distribution with $[\alpha, \beta]$, where $0 \leq \alpha < \beta \leq 1$.
- (8) Shortages are completely backordered.
- (9) A single product is considered.

3 Mathematical Model

We derive the cost involved in integrating the lot sizing policies between a vendor and a buyer. The ultimate form of JIT purchasing agreement should be adopted to minimize the total cost by implementing frequent small lots deliveries. Figure 1 depicts the behavior of inventory levels for both the vendor and the buyer. The annual integrated total cost consists of the vendor's annual total cost, and the buyer's annual total cost.

3.1 The Vendor's Total Cost per Unit Time

Figure 1 shows that in periodic delivery, the vendor does not stop producing until all demand is satisfied. For a given R , the values of T_{v1} and T_{v2} can be derived as

$$T_{v1} = R / p \tag{1}$$

and

$$T_{v2} = \frac{1}{\theta} \ln \left(\left(1 - \frac{p}{D} \right) \left(\exp \left(\frac{-R\theta}{p} \right) - 1 \right) + 1 \right) \tag{2}$$

respectively. The proof is attached in Appendix A.

For $T_{v1} + T_{v2} = T_v$, one has

$$T_{v1} + T_{v2} = \frac{R}{p} + \frac{1}{\theta} \ln \left(\left(1 - \frac{p}{D} \right) \left(\exp \left(\frac{-R\theta}{p} \right) - 1 \right) + 1 \right) \tag{3}$$

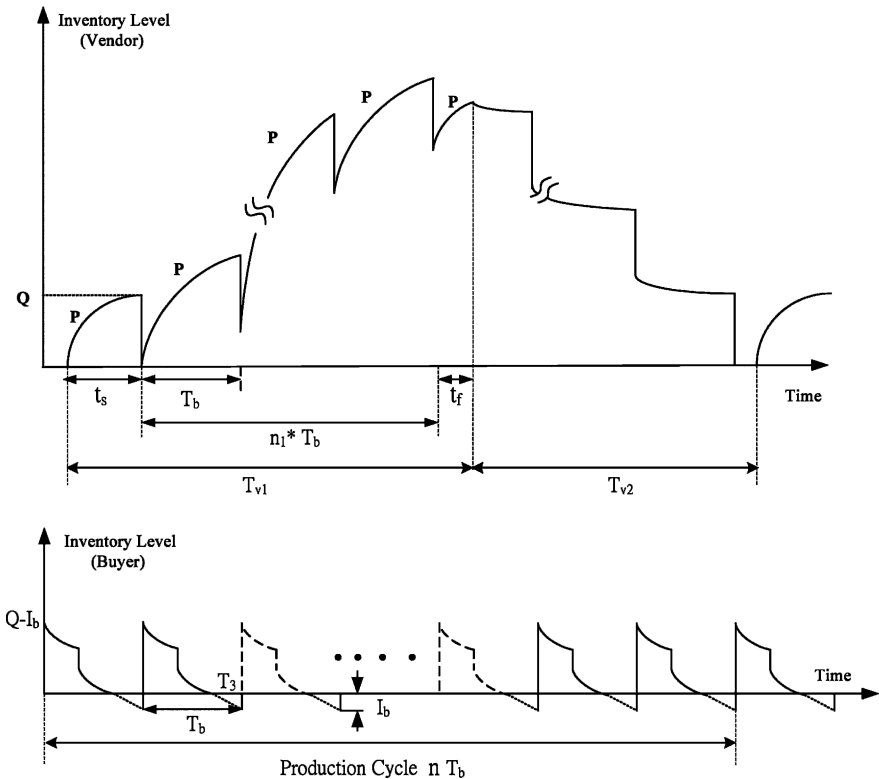


Fig. 1. Inventory level of vendor and buyer with customer demand

The vendor total inventory cost per unit time is depicted by the following formula:

Total cost = setting cost + delivery cost + holding cost + deteriorating cost

A carrying inventory can be derived as follows

$$\text{Carrying inventory} = \frac{\Delta I - D \cdot \Delta t}{\theta} = \frac{pT_{v1} - nQ}{\theta} = \frac{R - nQ}{\theta}$$

Hence, the holding cost per cycle is equal to $C_{vh} \left(\frac{R - nQ}{\theta} \right)$ and the deteriorating cost per cycle is equal to $C_{vd}(R - nQ)$. In addition to the setup cost and the delivery cost, the vendor's annual total cost is given by

$$TC_v = \frac{C_s}{T_v} + \frac{nC_d}{T_v} + \left(\frac{R - nQ}{T_v} \right) (C_{vh} + C_{vd}) \tag{4}$$

3.2 The Buyer's Total Cost per Unit Time

Figure 2 shows a lot size of Q units is replenished with an ordering cost of $\$K$ and a purchasing price of $\$c$ per unit. A fraction of each lot received is defective, with a known probability density function $f(p)$. The random variable p has a uniform distribution $[\alpha, \beta]$, where $0 \leq \alpha < \beta \leq 1$. A 100% screening process of the item is

Inventory Level

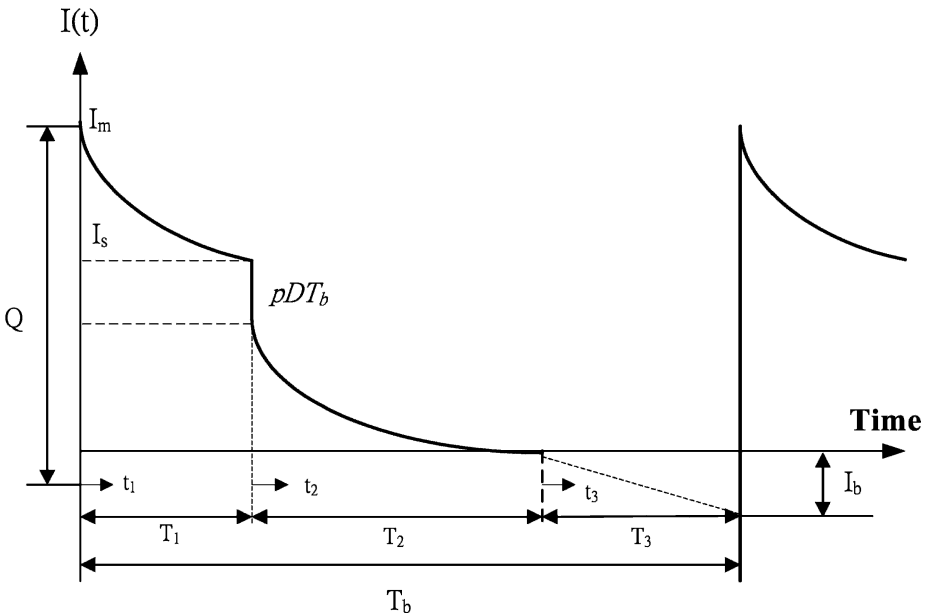


Fig. 2. Buyer's inventory system with backordering

conducted at a rate of X . The defective items are picked up in a single batch during the replenishment period T_1 . Shortages of stock are partial backlogged at the beginning of each period. The behaviour of the inventory system is illustrated in Figure 1, where T_b is the cycle length, pDT_b is the maximal number of defectives, and I_b is the total unit backordered.

The buyer's total cost per unit time, TC_b , is depicted as:

Total inventory cost = Ordering cost + Screening Cost + Deteriorating cost + Holding cost + Backordering cost.

For the inventory system depicted in figure 3, the carrying inventory within the time interval between T_1 and T_2 is

$$\frac{\Delta I - D \cdot \Delta t}{\theta} = \frac{1}{\theta}(Q - DT_b p - DT_b) \tag{5}$$

For $I_b = DT_3 = pDT_b$, the backlogged inventory during T_3 is equal to $\frac{1}{2}DT_b^2 p^2$.

Therefore, the total annual inventory cost is

$$TC_b = TC_b(T_b) = \frac{K}{T_b} + \left(c + x + d + \frac{h}{\theta} \right) \frac{Q}{T_b} - \left(d + \frac{h}{\theta} \right) (Dp + D) + \frac{bDp^2 T_b}{2} \tag{6}$$

The change in the inventory level during an infinitesimal time, dt , is a function of the deterioration rate θ , the demand rate D , and the inventory level $I(t)$. It is formulated as

$$\frac{dI_1(t)}{dt} + \theta I_1(t) = -D \quad 0 \leq t \leq T_1 \tag{7}$$

$$\frac{dI_2(t)}{dt} + \theta I_2(t) = -D \quad 0 \leq t \leq T_2 \tag{8}$$

$$\frac{dI_3(t)}{dt} = -D \quad 0 \leq t \leq T_3 \tag{9}$$

$I(t)$ is the inventory level at time t .

From the above differential equations, after adjusting for the constant of integration with various boundary conditions: $I_1(0) = I_s$, $I_2(0) = I_s - pDT$ and $I_3(0) = 0$, the differential equations become:

$$I_1(t_1) = \left[\left(I_s + \frac{D}{\theta} \right) \exp((T_1 - t_1)\theta) \right] - \frac{D}{\theta}, \quad 0 \leq t_1 \leq T_1 \tag{10}$$

$$I_2(t_2) = \left[\left(I_s - pDT + \frac{D}{\theta} \right) \exp(-\theta t_2) \right] - \frac{D}{\theta}, \quad 0 \leq t_2 \leq T_2 \tag{11}$$

and

$$I_3(t_3) = -Dt_3 \quad , \quad 0 \leq t_3 \leq T_3 \tag{12}$$

Since the defective items are independent of deterioration, they have a value equal to pDT_b . From figure 2, $I_1(T_1) = I_s = I_2(0) + pDT_b$, one has

$$I_s = pDT_b + \frac{D}{\theta} (\exp(\theta T_2) - 1) \tag{13}$$

$$I_b = DT_3 = pDT_b \tag{14}$$

and

$$I_1(0) = Q - pDT_b = \left(pDT_b + \frac{D}{\theta} (\exp(\theta T_2)) \right) (\exp(\theta T_1)) - \frac{D}{\theta} \tag{15}$$

The replenishment, Q , can be derived by substituting $T_1 = DT_b/X$ into Eq. (15). One has

$$Q = Q(T_b) = pDT_b + pDT_b \exp(\theta T_b D/X) + \frac{D}{\theta} \exp(\theta T_b(1-p)) - \frac{D}{\theta} \tag{16}$$

Expanding the exponential functions and neglecting the third and higher power of $\theta \cdot T$, Eq. (16) becomes:

$$Q = Q(T_b) = (Dp + D)T_b + \left(\frac{p\theta D^2}{X} + \frac{\theta D}{2}(1-p)^2 \right) T_b^2 + \left(\frac{p\theta^2 D^3}{2X^2} \right) T_b^3 \tag{17}$$

By substituting Eq. (17) into Eq. (6), one has

$$\begin{aligned} TC_b(T_b) = & \frac{K}{T_b} + \left(c + x + d + \frac{h}{\theta} \right) \left(Dp + D + \left(\frac{p\theta D^2}{X} + \frac{\theta D}{2}(1-p)^2 \right) T_b + \left(\frac{p\theta^2 D^3}{2X^2} \right) T_b^2 \right) \\ & - \left(d + \frac{h}{\theta} \right) \left(Dp + D \right) + \frac{bDp^2 T_b}{2} \end{aligned} \tag{18}$$

3.3 The Joint Total Cost Per Unit Time

For the vendor, by substituting $T_v = nT_b$ and Eq. (17) into Eq. (4), one has

$$\begin{aligned} TC_v(T_b) = & \left(\frac{R}{nT_b} - (Dp + D) - \left(\frac{p\theta D^2}{X} + \frac{\theta D}{2}(1-p)^2 \right) T_b - \left(\frac{p\theta^2 D^3}{2X^2} \right) T_b^2 \right) \left(\frac{C_{vh}}{\theta} + C_{vd} \right) \\ & + \frac{C_s}{nT_b} + \frac{C_d}{T_b} \end{aligned} \tag{19}$$

From Eq. (18) and Eq. (19), the total cost per unit time for both the vendor and buyer is: $JTC(T_b, n) = TC_v(T_b, n) + TC_b(T_b)$

For $\alpha \leq p \leq \beta$, one has

$$\begin{aligned}
 E[p] &= \frac{\alpha + \beta}{2} = \mu_1 \\
 E[p^2] &= \frac{\alpha^2 + \alpha\beta + \beta^2}{3} = \mu_2 \\
 E[(1-p)^2] &= \frac{\alpha^2 + \alpha\beta + \beta^2}{3} - (\alpha + \beta) + 1 = \mu_3
 \end{aligned}$$

The expected value of $JTC, EJTC$, is

$$\begin{aligned}
 EJTC(T_b) &= \left(\frac{R}{nT_b} - (D\mu_1 + D) - \left(\frac{\mu_1\theta D^2}{X} + \frac{\theta D}{2}\mu_3 \right) T_b - \left(\frac{\mu_1\theta^2 D^3}{2X^2} \right) T_b^2 \right) \left(\frac{C_{vh}}{\theta} + C_{vd} \right) \\
 &\quad + \frac{C_s}{nT_b} + \frac{C_d}{T_b} \\
 &\quad + \frac{K}{T_b} + \left(c + x + d + \frac{h}{\theta} \right) \left(D\mu_1 + D + \left(\frac{\theta D^2\mu_1}{X} + \frac{\theta D\mu_3}{2} \right) T_b + \left(\frac{\theta^2 D^3\mu_1}{2X^2} \right) T_b^2 \right) \\
 &\quad - \left(d + \frac{h}{\theta} \right) (D\mu_1 + D) + \frac{bDT_b\mu_2}{2}
 \end{aligned} \tag{20}$$

and from Eq. (3),

$$T_b = T_b(n) = \frac{1}{n} \left(\frac{R}{p} + \frac{1}{\theta} \ln \left[\left(1 - \frac{p}{D} \right) \left(\exp \left(\frac{-R\theta}{p} \right) - 1 \right) + 1 \right] \right) \tag{21}$$

3.4 Methodology and Solution Search

Our objective is to minimize the expected cost function.

$$\begin{aligned}
 &Min_{\alpha \leq p \leq \beta} EJTC(n) \\
 &s.t. \quad Q(T_b) > 0 \\
 &\quad T_b > 0 \\
 &\quad n \in N
 \end{aligned} \tag{22}$$

The problem is to determine the value of n that minimizes $EJTC$. Since the number of deliveries per cycle, n , is a discrete variable, the value of n can be derived as follows:

- Step 1. Input all the system parameters.
- Step 2. For a range of n -value, equate the first derivative of $EJTC$ with respect to T_b to zero. For each n , denote the resulting minimum value of T_b by $T_b(n)$.

Step 3. Derive the optimal value of n , denoted by $n^\#$. And, the value, n^* , is an integer in the vicinity of $n^\#$. The optimal value of must satisfy

$$EJTC(n^* - 1) \geq EJTC(n^*) \leq EJTC(n^* + 1) \tag{23}$$

Step 4. Using (17), the periodic delivery quantity, Q , can be solved.

4 Numerical Example

To illustrate the preceding theory, we compare our analysis with the example from Salameh and Jaber. The following data are assumed: $R=150000$ unit, $P=160000$ units/year, $C_s=300$ /cycle, $C_d=\$25$ /unit, $C_{vh}=\$2$ /unit/year, $C_{vd}=\$30$ /unit, $D=50000$ units/year, $K=\$100$ /cycle, $h=\$5$ /unit/year, $X=175200$ units/year, $x=\$0.5$ /unit, $b=\$10$ /unit/year, $c=\$25$ /unit, $d=\$30$ /unit, $s=\$50$ /unit. The item deteriorates at a constant rate with $\theta=0.01$. The percentage defective random variable, p , can take any value in the range $[\alpha, \beta]$ where $\alpha=0$, and $\beta=0.04$. It is assumed that p is uniformly distributed with its p.d.f.

$$f(p) = \begin{cases} 25, & 0 \leq p \leq 0.04, \\ 0 & \text{otherwise.} \end{cases}$$

Since $EJTC$ is a very complicated function due to high-power expression of the exponential function, a graphical representation showing the convexity of the $EJTC$ is given in Fig. 3. Following the above solution procedure, we compute the optimal value of n that minimizes Eq.(20) as $n^*=75$ ($n^\#=74.8$). Substituting $n^*=75$ into Eq.(21), the optimum values of T_b is 0.0396 year. From Eq.(17), the lot size Q^* is 2020 units. Therefore, the integrated total cost per year is \$1,194,719.

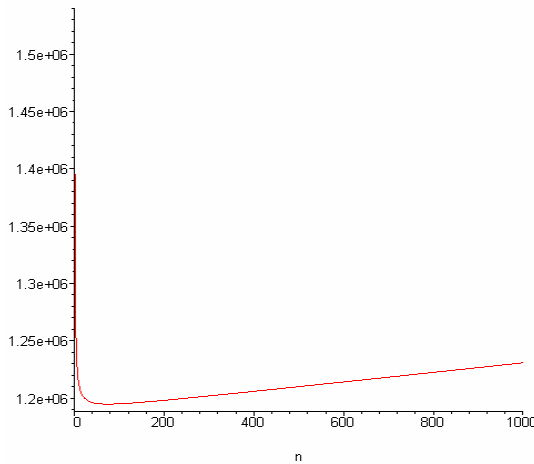


Fig. 3. Graphical representation of a convex $EJTC$ (when $n^*=75$)

Based on the numerical example, if the decision is made solely from the buyer's perspective, the optimal value of T_b that minimizes Eq.(18) is $T_b^*=0.0272$. From Eqs.(17) and (21), the optimal values of Q and n are $Q^*=1388$ and $n^*=109$. Substituting them into the buyer's expected annual cost and the vendor's expected annual cost, the total cost of the buyer and the vendor is \$1,195,172. Therefore, the integrated cost reduction is $(1,195,172-1,194,719)=453$. Note that the expected annual integrated total cost has an impressive cost-reduction as compared with an independent decision by the buyer.

5 Conclusions

This study has presented a deteriorating inventory model with unreliable process. The model extends the studies in [25] and considers a single-vendor single-buyer integrated two-echelon supply chain environment. Comparative studies in the example show the benefit of integration. The effect of deterioration should be considered even if it is small. We have shown in this study that the influence of imperfect quality, deterioration and complete backordering are significant. The management of an enterprise can select suppliers based on the defective percentage and the deterioration rate of the products supplied by each supplier.

References

1. Banerjee, A.: A joint economic lot size model for purchaser and vendor. *Decision Science* **17** (1985) 292-311
2. Cheng, T.C.E.: An economic order quantity model with demand-dependent unit production cost and imperfect production process. *IIE Transactions* **23**(1) (1991) 23-28
3. Chung, K.J., Hou, K.L.: An optimal production run time with imperfect production processes and allowable shortages. *Computers and Operations Research* **20** (2003) 483-490
4. Dave, U.: On a discrete-in-time order-level inventory model for deteriorating items. *Operational Research Quarterly* **30** (1979) 349-354
5. Elsayed, E.A., Terasi, C.: Analysis of inventory systems with deteriorating items. *International Journal of Production Research* **21** (1983) 449-460
6. Goyal, S.K.: Determination of optimum production quantity for a two-stage production system. *Operational Research Quarterly* **28** (1977) 865-870
7. Goyal, S.K.: A joint economic lot size model for purchaser and vendor: A comment. *Decision Science* **19** (1988) 236-241
8. Goyal S.K., Gupta, Y.P.: Integrated inventory models: the buyer-vendor coordination. *European Journal of Operational Research* **41** (1992) 261-269
9. Goyal, S.K.: A one-vendor multi-buyer integrated inventory model: a comment. *European Journal of Operational Research* **82** (1995) 209-210
10. Goyal, S.K., Gardenas-Barron, L.E.: Note on: economic production quantity model for items with imperfect quality – a practical approach. *International Journal of Production Economics* **77** (2002) 85-87
11. Goyal, S.K., Huang, C.K., Chen, H.K.: A simple integrated production policy of an imperfect item for vendor and buyer. *Production Planning & Control* **14**(7) (2003) 596-602

12. Ghare, P.M., Schrader, S.F.: A model for an exponentially decaying inventory. *Journal of Industrial Engineering* **14** (1963) 238-243
13. Ha, D., Kim, S.L.: Implementation of JIT purchasing: an integrated approach. *Production Planning & Control* **8**(2) (1997) 152-157
14. Hahm J., Yano, C.A.: The economic lot and delivery scheduling problem: the single item case. *International Journal of Production Economics* **28** (1992) 235-251
15. Hahm J., Yano, C.A.: The economic lot and delivery scheduling problem: the common cycle case. *IIE Transactions* **27** (1995) 113-125
16. Heng, K.J., Labban, J., Linn, R.L.: An order-level lot-size inventory model for deteriorating items with finite replenishment rate. *Computers & Industrial Engineering* **20** (1991) 187-197
17. Hill, R.M.: The single-vendor single-buyer integrated production-inventory model with a generalized policy. *European Journal of Operational Research* **97** (1997) 493-499
18. Hill, R.M.: The optimal production and shipment policy for the single-vendor single-buyer integrated production inventory problem. *International Journal of Production Research* **37** (1999) 2463-2475
19. Hoque, M.A., Goyal, S.K.: An optimal policy for a single-vendor single-buyer integrated production-inventory system with capacity constraint of transport equipment. *International Journal of Production Economics* **65** (2000) 305-315
20. Kang, S., Kim, I.: A study on the price and production level of the deteriorating inventory system. *International Journal of Production Research* **21** (1983) 449-460
21. Lu, L.: A one-vendor multi-buyer integrated inventory model. *European Journal of Operational Research* **81** (1995) 312-323
22. Porteus, E.L.: Optimal lot sizing, process quality improvement and setup cost reduction. *Operations Research* **34**(1) (1986) 37-144
23. Raafat, F., Wolfe, P.M., Eldin, H.K.: An inventory model for deteriorating items. *Computers & Industrial Engineering* **20** (1991) 89-94
24. Rosenblatt, M.J., Lee, H.L.: Economic production cycles with imperfect production process. *IIE Transaction* **18** (1986) 48-55
25. Salameh, M.K., Jaber, M.Y.: Economic order quantity model for items with imperfect quality. *International Journal of Production Economics* **64** (2000) 59-64
26. Schwaller, R.L.: EOQ under inspection costs. *Production and Inventory Management* **29**(3) (1988) 22
27. Viswanathan, S.: Optimal strategy for the integrated vendor-buyer inventory model. *European Journal of Operational Research* **105** (1998) 38-42
28. Wee, H.M., Yu, J., Chen, M.C.: Optimal inventory model for items with imperfect quality and shortage backordering. *Omega*. (2005) In press (available online)
29. Yang, P.C., Wee, H.M.: A single-vendor and multiple-buyers production-inventory policy for a deteriorating item. *European Journal of Operational Research* **43** (2002) 570-581
30. Zhang X., Gerchak, Y.: Joint lot sizing and inspection policy in an EOQ model with random yield. *IIE Transaction* **22**(1) (1990) 41

Appendix A

From Fig. 2, the vendor's production interval, T_{v1} , and the production interval, T_{v2} , can be denoted by

$$T_{v1} = t_s + n_1 \cdot T_b + t_f \quad \text{and} \quad T_v = T_{v1} + T_{v2} = n \cdot T_b \quad (\text{A1})$$

It is observed that the total inventory level at T_{v1} (the production stage) is the same as total inventory level at T_{v2} (the non-production stage). One has

$$I(T_{v1}) = \left(\frac{p-D}{\theta} \right) (1 - \exp(-T_{v1}\theta)) = \left(\frac{D}{\theta} \right) (\exp(T_{v2}\theta) - 1). \quad (\text{A2})$$

Substituting $T_1 = R/P$ into (A2), the values of T_2 can be derived as

$$T_{v2} = \frac{1}{\theta} \ln \left(\left(1 - \frac{p}{D} \right) \left(\exp \left(\frac{-R\theta}{p} \right) - 1 \right) + 1 \right) \quad (\text{A3})$$

A Clustering Algorithm Using the Ordered Weight Sum of Self-Organizing Feature Maps

Jong-Sub Lee¹ and Maing-Kyu Kang²

¹Department of Technical Management Information Systems, University of Woosong,
Daejeon, South Korea
ljs@wsu.ac.kr

²Department of Information & Industrial Engineering, University of Hanyang,
Ansan, South Korea
dockang@hanyang.ac.kr

Abstract. Clustering is to group similar objects into clusters. Until now there are a lot of approaches using Self-Organizing Feature Maps(SOFMs). But they have problems with a small output-layer nodes and initial weight. This paper suggests one-dimensional output-layer nodes in SOFMs. The number of output-layer nodes is more than those of clusters intended to find and the order of output-layer nodes is ascending in the sum of the output-layer node's weight. We can find input data in SOFMs output node and classify input data in output nodes using the Euclidean Distance. The suggested algorithm was tested on well-known IRIS data and machine-part incidence matrix. The results of this computational study demonstrate the superiority of the suggested algorithm.

1 Introduction

Clustering deals with data corresponding to processes of clusters. The predicament of clustering is the input of n data of multi dimension and dividing it into k cluster having similar features. When compared with common data, data in a cluster has greater similarities rather than the differences. The measurement of the similarity is calculated by the Euclidean Distance Method based on the Attribute Value of Data. The shorter the size of Euclidean distance, the higher the resemblance.

Clustering Algorithms can be divided into two categories; hence Hierarchical Algorithms and Partition Algorithms. Recently, there has been a dynamic study of clustering algorithms utilizing neural network and fuzzy-neural networks. The hierarchical algorithm, which is postulated by Mangiameli *et al*[10], depicts that there is single linkage clustering which measures the similarity by the minimum distance between two clusters, complete linkage clustering which measures the similarity by the maximum distance between two clusters, group-average clustering which measures the similarity by the average distance between the two clusters, and finally consistent with Ward's Hierarchical Clustering measure the similarity by density between two clusters. Hierarchical Algorithms has a defect in that it cannot complete the problems caused by an inappropriate merge.

In a broader way, while we consider about the Partition Algorithms, it can be noted that such algorithms formulate partition of the data and form clusters so that data in a

cluster is more similar than other clusters. Partition Algorithms can be characteristically classified into k -Means Algorithm and ISODATA Algorithm. To deal with k -Means Algorithm, this will incorporate the data of each cluster which is able to rearrange as the algorithms is repeated into the direction minimizing the distance difference between each data and central value of each cluster. This conquered the disadvantages of hierarchical algorithms, which was unable to overcome the inappropriate merge occurring in the early stage. ISODATA Algorithm start with k centurms but the number of clusters is not necessarily k . Despite this aspect, the number of clusters can be flexibly changed during algorithm performing. ISODATA algorithm complete k -Means algorithm' defect of having fixed numbers of clusters[5].

In the year of 1989, two professionals called Huntsberger and Ajjimarangsee[6] postulated a clustering algorithm modified in parameters such as learning rate and vicinity rate and slightly modified Kohonen[8]'s learning method. In 1993 another personality called Pal *et al*[11] illustrated a lose function method which provided the connecting weight of the distance between input data and output node by early connection strength, its also designated the Competitive learning neural network Algorithms that minimize the Lose function.

The fuzzy concept of neural network was divulged by two individuals named Tasao *et al*[13] and Karayiannis[7]. Particularly with Karayiannis[7] he formulated an algorithm that minimizes the connecting weight sum of square Euclidean distance between FALVQ(Fuzzy Algorithm for Learning Vector Quantization) inputting data and connecting weight of LVQ(Learning Vector Quantization).

In 2000 another person called Kusiak[9] defined the clustering problem as an NP-Complete problem. This illustrates where the number of machines in a cluster is higher, the computational complexity is exponentially increased and time consuming. To overcome such tribulations, we prefer Heuristic algorithms to Optimization algorithms.

In this erudition, the attribute of the connecting weight modifying the form into a similar one to the inputting data is consumed, when the Self-Organization neural network is in a progress of unsupervised learning with input data of Anderson's IRIS data and machine-part matrix data. The attributes of this cram depict the numbers of one dimension output nodes, learning rate and the boundary of adjacency, etc. The suggested algorithm creates a process of learning using these parameters, and groups depending on the dimensions of Connecting Weight Distance between i and j output nodes. As the result this algorithm elaborates how to reduce the numbers of misclassification.

2 SOFMs Neural Network

SOFM is a Competitive Learning Neural Network model which is explained by Kohonen[8]. Fig. 1 illustrates that SOFM includes two layers, hence an input layer formed with m input nodes and an output layer formed with n output nodes.

As the Input layer receives input data, mapping is performed at the output layer. The output layer uses either a one-dimensional structure or a two-dimensional structure. We can set the output node to either a bigger number than that of input data or a random number k chosen by users so that the input data can be spread on other output nodes. At each node of output, mapping is performed with the input data.

Every node of input layer and output are connected and there will be a connecting line between output node i ($1 \leq i \leq n$) and input node j ($1 \leq j \leq m$), having a Connecting

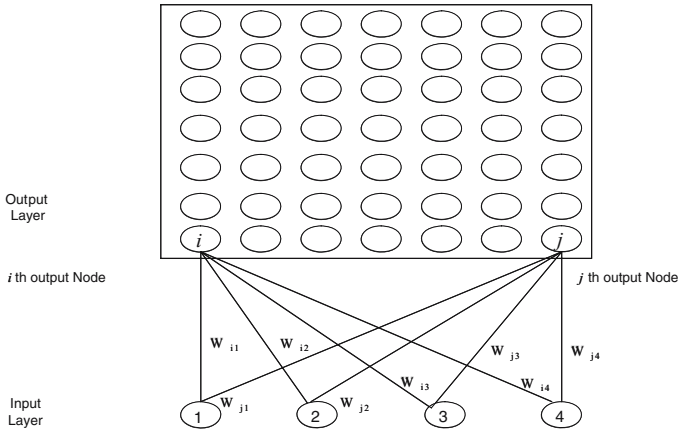


Fig. 1. General structure of SOM

Weight W_{ij} . Connecting weight is a real number between 0 and 1, which is randomly presented at the initial stage but can be changed by the input data. Corresponding with each input data, winner node i^* is selected, which is the most similar output node. As formula (1), among the distance D_i calculated between input data X and connecting weight W_i , the lowest output node i^* is selected

$$D_i = \sqrt{(x_1 - w_{i1})^2 + (x_2 - w_{i2})^2 + \dots + (x_m - w_{im})^2} \quad i = 1, \dots, n \tag{1}$$

Connecting weight W_{i^*} of output node i^* which is given in as formula (2), will establish the nearest connecting weight to input data X among n output node. The node located on the front and the back of winner node i^* is called the Neighbor Node and Neighborhood $N_{i^*}(\delta)$ is a congress of neighbor node separated as δ from winner node i^* . Winner node is specified as “#” in Fig. 2. While we refer to the other output node as “*”, the neighbor node is referred with the case of which the Radius is $\delta=0, 1, 2$ by using the Rectangular Grid.

$$|X - W_{i^*}| = \min |X - W_i| \quad i = 1, \dots, n \tag{2}$$

Coherence with neighborhood as $N_{i^*}(\delta)$, the connecting weight of self-organization neural network decreases the neighbor range and the learning ratio with regulating the connecting weight until the neighbor range becomes the same as the winner node itself as formula (3). The learning ratio $\alpha(t)$ is a ratio which is used to control the difference between input data and existing connecting weight in accordance with flow of time t . This is a real number between 0 and 1 and progressively decreases as the learning proceeds. Generally, the learning cannot be achieved accurately while the learning ratio is too high, and it takes too long when its too low[9].

Here $w(old)_{ij}$ represents the link-weight before adjustment and $w(new)_{ij}$ represents the link weight after adjustment.

$$w(new)_{ij} = w(old)_{ij} + \alpha(x_i - w(old)_{ij}) \quad i \in N_{i^*}(\delta) \quad j = 1, \dots, m \tag{3}$$

$N_{i^*}(\delta)$ is neighborset with δ far from i^*

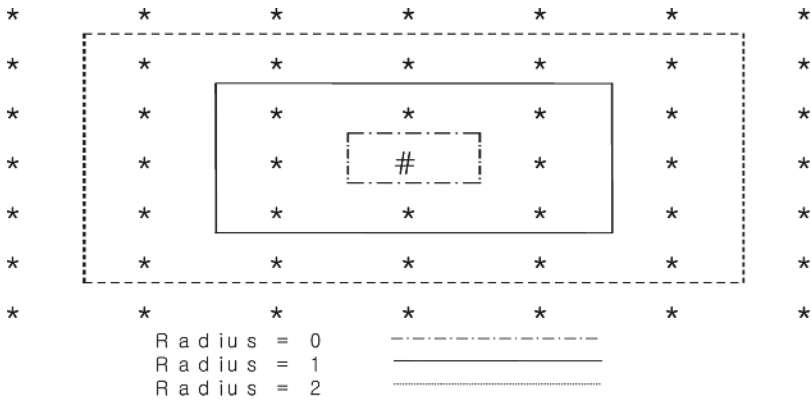


Fig. 2. The neighborhood using rectangular grid in two dimensions

Learning algorithm is summarized as follows.

1. Initialize link weight with random number between 0 and 1. Decide the range of neighbor and the learning rate.
2. Input one input vector in input layer.
3. Calculate the distance D_i between the i th input vector and link weight vector of formula (1).
4. Choose one winner node.
5. Adjust link weight according to learning rule of formula (3).
6. Downsize the range of neighbor and learning rate. Repeat the process from step 2 to 5 until the range of neighbor becomes winner node itself.

3 Suggestion

Taking into account at first level, the SOFM given by Kohonen has a few tribulations which are caused by the change of solution and an extensive amount of learning time. The suggested structure of SOFM formed with a one-dimensional linear output layer gets a connecting weight sum at each output node based on a randomly given connecting weight at first level. Anchored in this, it lines up in an ascending order. Matching input data to the consisted output node and transforming output node to a similar form to input node through learning, it forms a group. Exceptionally, when the number of output node is set higher than the number of input data, it enables the distribution of the output node to be spread in similar form as the distribution of input data. During the learning process, at the point of neighbor range's being half at the first level, the connecting weight of output node that had not been winner node is not regulated as formula (3). In accordance with the same situation, when learning processed input data is lined up on the output node which has a similar order, the required group can be formed by linear separating the highest point of the connecting weight among output nodes. This disentangles the defect which can occur in learning process, and apparently acts upon the process of forming a group after learning.

Suggested clustering algorithm settles on the parameter such as a number of output node, learning ratio, and neighbor range. Smaller the number of output node, shorter the learning time. Least number of output node enables the input data to adequately spread on the output node. According to practical method, the best solution is shown by how to trounce the drawbacks which are possibly caused from the learning process; it would be gained if the number of output node is set to more than twice of the number of input data. As the learning ratio gets higher, the learning time decreases. At first level, set every output mode as neighbor, diminish the range as time passes, and stop the algorithm when the neighbor range is itself. Set the initial learning ratio, $\alpha(0)$, to 0.4 which practically offers good results, and set the initial neighbor range as the same number as that of output node.

With the assistance of Euclidean distance, separate the output node i , on which each input data is mapped from the distance between the connecting weight $i+1$. Where i and $i+1$ represent the number of neighboring output node. That is why the connecting weight consists of a similar form of probability distribution function; the mapped neighboring output node's connecting weight is used as Euclidean distance. So the neighboring output node's connecting weight simply needs to be considered.

Suggested Clustering algorithm process is as follows,

1. Initialize the structure of SOFM (output node type, number of input or output node)
2. Initialize the weight on each connecting line, and set an initial learning ratio $\alpha(0)$ and learning ratio function $\alpha(t)$, and an initial neighboring range.
3. Gain the sum of connecting ratio at each output node. Depends on the volume of connecting weight sum, arrange the output node with an ascending order.
4. As formula (1), refers to each input data, calculate the distance between the connecting weight of output node and input data and decide the winner node which is the shortest node.
5. Regulate the connecting ratio of neighbor node located at a regular range from the winner node.
6. Decrease the neighbor range as 1 and learning ratio as $\alpha(t)=(1-t/4950)\times\alpha(t-1)$. Finally, repeat Step 4 to 5 until the neighbor range becomes winner node itself or learning ratio to be 0. Exclude the tip that the neighbor range becomes half of the initial neighbor range, hence any disastrous output node becomes a winner node and thereafter learning is not required.
7. Map each input data on the nearest output node.
8. Calculate the connecting weight distance $WD(i, j)$ between the neighboring output node i and j of which the input data is matched.
9. At the point of an output node with linear structure, selecting $k-1$ (output node range) which has the greatest difference of connecting weight distance $WD(i, j)$ is able to form k group.

4 Numerical Example

In this premise, we recycled the postulations presented by Anderson's IRIS data: [1] and machine-part matrix: [2],[3],[4],[9],[12]. Anderson's IRIS data, especially,

consists of 150 samples of data which have parameters of 4 dimensions (Petal Width, Petal Length, Sepal Width, and Sepal Length). The three clusters (Iris Setosa, Iris Versicolour, Iris Verginica) consist of 50 pieces of data each. It is generally noted that IRIS data, as a clustering algorithm which devours unsupervised learning, is known to fabricate about 15 to 17 misclassifications[11].

Following are some distinctive applications of suggested clustering algorithms to IRIS data used for this study.

1. The structure of SOFM used for this example is depicted as is in the subsequent case. The structure of the output node is linear and the number of the input nodes and the output nodes are 4, 300 each. In this situation, the structure of the output nodes is also linear, and the number of the output doubles the input data. Hence, the number of input nodes is 4.

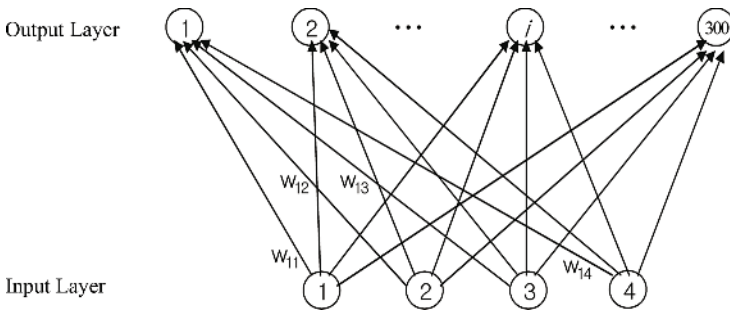


Fig. 3. Suggesting structure of SOFMs

2. The weight of the first output node is $W_1=\{0.5882, 0.2500, 0.2200, 0.1872\}$, the weight of the second output node is $W_2=\{0.9232, 0.4950, 0.8613, 0.7534\}$, the weight of nineteenth output node is $W_{19}=\{0.0145, 0.2887, 0.0584, 0.0835\}$ and the weight of 300th output node is $W_{300}=\{0.7083, 0.8935, 0.8302, 0.0759\}$. Set the initial learning ratio as 0.4, and the learning function is $\alpha(t)=(1-t/4950)\times\alpha(t-1)$.
3. The sum of the weight of the output nodes: the first: 1.245, the second: 2.5468, the 19th :0.4453, the 300th: 2.5082. The order by size is this: 19, 48, ..., 44, 89.
4. When calculating the distance W_{19} , the weight of the first output to the first input data as formula (1), it produces 0.6902, 1.3757, 1.6861, And the shortest output node, W_{19} , is called the winner node.
5. Adjust the weight of all output nodes within 300 radiuses as formula (3).
6. Reduce the neighbor rate one by 1, and in the case when t is zero, reduce the initial learning ratio $\alpha(0)$ to 0.4, and in the case when t is 1, reduce it as $0.4\times(1-(1/4950))$ and repeat the Step 4 to 5 until the neighbor rate becomes the winner node itself.
7. When mapping each data to the nearest output node, it is as Table 1.
8. When calculating the distance of each output node and the weight distance, the results are as follow: $WD(1, 5)=0.011$, $WD(5,8)=0.0074$, $WD(8,10)=0.1738$, $WD(10,12)=0.0286$, ... , $WD(56, 75)=2.8132$, ... , $WD(182, 191)=0.4176$, ... , $WD(299, 300)=0.0139$.

Table 1. The Data in the output nodes

No ¹	Data ²	No	Data	No	Data	No	Data	No	Data
1	6,11	41	14,43	100	90	169	59	241	133
	15,16	42	39	102	63	170	55	245	116,149
	17,19	44	4,9,	105	83,93	175	51,77,87	250	142,146
	33,34		13	111	68	176	53,57	253	137
5	37,49	47	2,10	114	107	182	78	260	105
8	20,45		46	115	91	191	73,134	261	101
	47	48	42	116	95	196	124,127	262	141,145
10	21	49	35	119	100	202	128,139	263	113
12	22,32	50	30	121	89	204	71	264	140
15	28	52	31	125	97	205	120	266	125
18	18	54	26	126	85,96	206	84	267	109
		56	25	128	56,67	211	150	271	121
19	1,5, 29,44	75	99	134	62	213	122	272	144
		77	58	136	72	214	102,114	280	103
21	24,40	78	94	144	79,86		143,147	283	130
22	27,41	80	61	147	98	218	115	284	126
23	8,38	83	80	150	69,88	219	135	286	110
27	50	85	82	151	74	227	104	293	131
29	12	89	81	153	92	232	138,148	294	108
31	36	93	65	154	64	236	112	299	136
32	23	96	70	157	75	237	117	300	106,118
34	7	97	54	166	52,76	238	111		119,123
39	3	98	60	167	66	240	129		132
40	48								

Table 2. The Data in 3 Groups

No of Group	No of Input Data
1	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50
2	51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 72, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 107
3	71, 73, 84, 101, 102, 103, 104, 105, 106, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150

9. The output node and the section of output node that have the longest weight distance are 56 and 75 and the next are 182 and 191. Each value is $WD(56, 75)=2.8132$, $WD(182, 191)=0.4176$. Therefore, the 3 groups are same as table 2.

¹ No of Output Node.

² No of Input Data.

5 Result

The suggested revision is an assessment on IRIS data[1], referred by existing Studies and machine-part matrix data. As Table 2 shows, the misclassified data appear one in group 2, three in group 3 and the total is four. This analysis created a better consequence than the existing algorithms that appears in Table 3.

The recommended structure of SOFM of the output node is one-dimensional linear, nevertheless as the initial arbitrarily set weight; the sum of weight of the output node was not able to line up from left to right by its magnitude. To make it acceptable, we lined them up setting up the sum of weight as a standard from left to right in ascending order. In addition to this, for fixing on the number of output nodes which is probable to get the nearest optimizing quotient of IRIS data and machine-part matrix, we have carried out the experiment by putting the number of output nodes from 3 as a start, and increased it to quadruple input data. As a result, we found that the results under the situation of 3 output and less than the double of input data were insignificantly different from time to time; however, the result under the situation of more than two times of input data were the same.

The initial neighbor range is set as 300 in the radius, so as to enable amending the weight of all output nodes. Once the learning process passed over half of the initial neighbor range, the output nodes failed to become winner nodes, hence cease adjusting the weight.

Table 3 shows the result of measure up to the quotients of the suggested clustering algorithms to Anderson's IRIS data and of existing algorithms.

It can be widely known that the suggested clustering algorithms accomplished the quotient with 4 misclassifications, which is a enhanced result from Pal *et al*[11]'s 17 misclassifications and Karayiannis[7]'s 15 misclassifications.

According to Pal *et al*[11], the existing clustering algorithms that use unsupervised learning produce at least 15 to 17 misclassifications. The suggested clustering algorithms produce only 4 misclassifications which is lesser than presently existing algorithms.

The suggested clustering algorithms used the same parameter to solve the machine-part grouping problem which is well known in manufacturing field. When applied to IRIS data that has a value of 4 dimensional real number, setting the initial learning ratio as 0.4, and the learning function $(1-t \times (1/4950))$. The machine-part grouping problem consists of the machine-part matrix, which has no exceptional elements.

Table 3. The number of error Comparison for Anderson's IRIS Data

Source of Problem	Source of Algorithms	No of Error
Anderson's IRIS Data Set	Suggested algorithm	4
	Karayiannis[7]	15
	Pal <i>et al</i> [11]	17

In the initial phase of machine-part grouping, Table 4 convinced the optimizing number of group and misclassifications that may occur during the grouping. The suggested clustering algorithm forms the machine cells, utilizing an independent machine-part matrix, and indicates the number of machine cells and misclassifications that will occur during the process, as Table 4. The suggested clustering algorithms indicate the optimizing number of machine cells and the minimum number of classifications, 0.

Table 4. The number of error Comparison for Machine-Part Incidence Matrix

Size ³	Source of Problems	No of Group		No of Error	
		Optimal	Suggested Algorithm	Optimal	Suggested Algorithm
4x5	Kusiak[9]	2	2	0	0
10x15	Chan <i>et al</i> [2]	3	3	0	0
10x20	Srinivasan <i>et al</i> [12]	4	4	0	0
24x40	Chandraseharan <i>et al</i> [4]	7	7	0	0
40x100	Chandraseharan <i>et al</i> [3]	10	10	0	0

6 Conclusion

This revise contributes an effective clustering algorithms classifying the IRIS data, and forming the machine-part groups. The features portrayed here are the structure of SOFM studying parameter. The structure of SOFM suggested in a one-dimension is linear. The number of output nodes is set as twice as the number of input node.

We endowed the weight to the output node voluntarily, and arranged them as the size of the sum of weight in ascending order. Once the learning is in process, and the neighbor range arrives at the point of half of the initial neighbor node, the weight of output node that has failed to be the winner node stops adjusting the weight. Analogous with output node the input data learning is completed and takes its turns, it is possible to form a group by linear dividing the point that has the largest difference of weight between output nodes.

According to the experienced method, to set the number of output node to more than twice the number of input data is the best way to achieve the best quotient that is able to overcome the defects that might occur during the learning process. In the neighbor arrangement, it set all output nodes as its neighbor, and as time goes by, it reduces the range, and finally when it comes to be its own neighbor, the algorithms stops.

A well recognized method called IRIS data and machine-part matrix is used in this premise. As the result of research on IRIS data, we achieved better quotient with only 4 misclassifications. But the normal way of the existing clustering algorithms utilizing unsupervised learning produces 15 to 17 misclassifications. So, broadly saying, the suggested algorithms do not use a complex operation, hence the suggested algorithms perform more flexibly and feasibly in real time applications.

³ No of Machinex No of Part.

References

1. Anderson, E.: The IRIS's of the Gaspé Peninsula, *Bull. Amer. IRIS Soc.*, 59 (1939) 2-5
2. Chan, H. M., Milner, D. A.: Direct clustering algorithm for group formation in cellular manufacturing, *Journal of Manufacturing Systems*, 1(1) (1982) 65-75
3. Chandrasekharan, M. P., Rajagopalan, R.: ZODIAC: An algorithm for concurrent format of part-families and machine-cells, *International Journal of Production Research*, 25(6) (1987) 835-850
4. Chandrasekharan, M. P., Rajagopalan, R.: Groupability: An analysis of the properties of binary data matrices for group technology, *International Journal of Production Research*, 27(6) (1989) 1035-1052
5. Everitt, B. S., Landau, S., Leese, M.: *Cluster Analysis*, Edward Arnold, London (2001)
6. Huntsberger, T. L., Ajjimarangsee, P.: Parallel Self-Organizing Feature Maps for Unsupervised Pattern Recognition, *International Journal of General Systems*, 16(4) (1990) 357-372
7. Karayiannis, N. B.: A Methodology for Constructing Fuzzy Algorithms for Learning Vector Quantization, *IEEE Trans. Neural Networks*, 8(3) (1997) 505-518
8. Kohonen, T.: *Self-Organizing Maps*, Springer, Berlin (1997)
9. Kusiak, A.: *Computational Intelligence in Design and Manufacturing*, John Wiley & Sons, New York (2000)
10. Mangiameli, P., Chen, S. K., West, D.: A Comparison of SOM Neural Network and Hierarchical Clustering Methods, *European Journal of Operational Research*, 93(2) (1996) 402-407
11. Pal, N. N., Bezdek, J. C., Tasao, E. C. -K.: Generalized Clustering Networks and Kohonen's Self-Organizing Scheme, *IEEE Trans. Neural Networks*, 4(4) (1993) 549-551
12. Sirmivasan, G., Narendran, T.T., Mahadevan, B.: An assignment model for the part families problem in group technology, *International Journal of Production Research*, 28(1) (1990) 145-152
13. Tasao, E. C. -K., Bezdek, J. C., Pal, N. N.: Fuzzy Kohonen Clustering Networks, *Pattern Recognition*, 27(5) (1994) 754-757

Global Optimization of the Scenario Generation and Portfolio Selection Problems

Panos Parpas and Berç Rustem

Department of Computing, Imperial College, London SW7 2AZ

Abstract. We consider the global optimization of two problems arising from financial applications. The first problem originates from the portfolio selection problem when high-order moments are taken into account. The second issue we address is the problem of scenario generation. Both problems are non-convex, large-scale, and highly relevant in financial engineering. For the two problems we consider, we apply a new stochastic global optimization algorithm that has been developed specifically for this class of problems. The algorithm is an extension to the constrained case of the so called diffusion algorithm. We discuss how a financial planning model (of realistic size) can be solved to global optimality using a stochastic algorithm. Initial numerical results are given that show the feasibility of the proposed approach.

1 Introduction

We consider the global optimization of two problems arising from financial applications. The first problem originates from the portfolio selection problem when high-order moments are taken into account. This model is an extension of the celebrated mean-variance model of Markowitz[1, 2]. The inclusion of higher order moments has been proposed as one possible augmentation to the model in order to make it more applicable. The applicability of the model can be broadened by relaxing one of its major assumptions, i.e. that the rate of returns are normal. The second issue we address is the problem of scenario generation i.e. the description of the uncertainties used in the portfolio selection problem. Both problems are non-convex, large-scale, and highly relevant in financial engineering.

Given the numerical and theoretical challenges presented by these models we only consider the “vanilla” versions of the two problems. In particular, we focus on a single period model where the decision maker (DM) provides as input preferences with respect to mean, variance, skewness, and possibly kurtosis of the portfolio. Using these four parameters we then formulate the multi-criteria optimization problem as a standard nonlinear programming problem. This version of the decision model is a non-convex linearly constrained problem.

Before we can solve the portfolio selection problem we need to describe the uncertainties regarding the returns of the risky assets. In particular we need to specify: (1) the possible states of the world and (2) the probability of each state. A common approach to this modeling problem is the method of matching moments (see e.g. [3, 4, 5]). The first step in this approach is to use the historical

data in order to estimate the moments (in this paper we consider the first four central moments i.e. mean, variance, skewness, and kurtosis). The second step is to compute a discrete distribution with the same statistical properties as the ones calculated in the previous step. Given that our interest is on real-world applications we recognize that there may not always be a distribution that matches the calculated statistical properties. For this reason we formulate the problem as a least squares problem [3, 4]. The rationale behind this formulation is that we try to calculate a description of the uncertainty that matches our beliefs as well as possible. The scenario generation problem also has a non-convex objective function, and is linearly constrained.

For the two problems described above we apply a new stochastic global optimization algorithm that has been developed specifically for this class of problems. The algorithm is described in [6] (see also section 4). It is an extension to the constrained case of the so called diffusion algorithm [7, 8, 9, 10]. The method follows the trajectory of an appropriately defined Stochastic Differential Equation (SDE). Feasibility of the trajectory is achieved by projecting its dynamics onto the set defined by the linear equality constraints. A barrier term is used for the purpose of forcing the trajectory to stay within any bound constraints (e.g. positivity of the probabilities, or bounds on how much of each asset to own).

The purpose of this paper is to show that stochastic optimization algorithms can be used to solve realistic financial planning problems. A review of applications of global optimization to portfolio selection problems appeared in [11]. A deterministic global optimization algorithm for a multi-period model appeared in [12]. This paper extends and complements the methods mentioned above in the sense that we describe a complete framework for the solution of a realistic financial model. The type of models we consider, due to the large number of variables, cannot be solved by deterministic algorithms. Consequently, practitioners are left with two options: solve a simpler, but less relevant model, or use a heuristic algorithm (e.g. tabu-search or evolutionary algorithms). The approach proposed in this paper lies somewhere in the middle. The proposed algorithm belongs to the simulated-annealing family of algorithms, and it has been shown in [6] that it converges to the global optimum (in a probabilistic sense). Moreover, the computational experience reported in [6] seems to indicate that the method is robust (in terms of finding the global optimum) and reliable. We believe that such an approach will be useful in many practical applications. Admittedly the models (especially the portfolio selection problem) are rather simplistic. Given the theoretical and computational difficulties involved with such models it is important to consider the simplified version of the problem in the hope that this approach will shed more light to the general case. Moreover, to the authors' knowledge this is the first paper to address, in a holistic manner, the global optimization of the moment problem and the portfolio selection problem with higher order moments.

The rest of the paper is structured as follows: in section 2 we describe the scenario generation problem. While there are many ways to generate scenario trees for stochastic programming problems, we will focus on the moment matching

approach. The interested reader is referred to [3] for a review of other methods. Also in this section we discuss the importance of arbitrage opportunities; we describe how we dealt with this requirement of financial models in our implementation. In section 3 we discuss the portfolio selection problem. A model with a non-convex objective and linear constraints is proposed as a simple extension to the classical Markowitz model. The non-convexities in the model arise from the inclusion of higher order moments. The model considered here relaxes the normality assumption of the classical model, the reader is referred to [13] for a more complete overview of non-convex optimization problems in financial applications. In section 4 we describe an algorithm for the solution of the two models described above. For a full theoretical treatment of the algorithm we refer the interested reader to [6]. In section 5 we present some initial numerical experiments. We study how difficult (in terms of computation time) it is to compute an arbitrage free scenario tree. We also study how the global optimum changes as we vary the parameters of the model. To illustrate the effect of the parameters we present some 3-dimensional plots of efficient frontiers, the analogs of the classical Markowitz efficient frontiers.

2 Scenario Generation

From its inception Stochastic Programming (SP) has found several diverse applications as an effective paradigm for modeling decisions under uncertainty. The focus of initial research was on developing effective algorithms for models of realistic size. An area that has only recently received attention is on methods to represent the uncertainties of the decision problem.

A review of available methods to generate meaningful descriptions of the uncertainties from data can be found in [3]. We will use a least squares formulation (see e.g. [3, 4]). It is motivated by the practical concern that the moments, given as input, may be inconsistent. Consequently the best one can do is to find a distribution that fits the available data as well as possible. It is further assumed that the distribution is discrete. Under these assumptions the problem can be written as:

$$\begin{aligned} \min_{\omega, p} \quad & \sum_{i=1}^n \left(\sum_{j=1}^k p_j m_i(\omega_j) - \mu_i \right)^2 \\ \text{s.t.} \quad & \sum_{j=1}^k p_j = 1 \quad p_j \geq 0 \quad j = 1, \dots, k \end{aligned}$$

Where μ_i represent the statistical properties of interest, and $m_i(\cdot)$ is the associated ‘moment’ function. For example, if μ_i is the target mean for the i^{th} asset then $m_i(\omega_j) = \omega_j^i$ i.e. the j^{th} realization of the i^{th} asset. Numerical experiments using this approach for a multistage model, were reported in [4] (without arbitrage considerations). Other methods such as maximum entropy [14], and semidefinite programming [15], while they enjoy strong theoretical properties

they cannot be used when the data of the problem are inconsistent. A disadvantage of the least squares model is that it is highly non-convex which makes it very difficult to handle numerically. These considerations have led to the development of the algorithm described in section 4 (see also [6]) that can efficiently compute global optima for problems in this class.

When using scenario trees for financial planning problems it becomes necessary to address the issue of arbitrage opportunities [4, 16]. An arbitrage opportunity is a self-financing trading strategy that generates a strictly positive cash flow in at least one state and whose payoffs are nonnegative in all other states. In other words it is possible to get something for nothing. In our implementation we eliminate arbitrage opportunities by computing a sufficient set of states so that the resulting scenario tree has the arbitrage free property. This is achieved by a simple two step process. In the first step we generate random rates of returns, these are sampled by a uniform distribution. We then test for arbitrage by solving the system:

$$x_0^i = e^{-r} \sum_{j=1}^m x_j^i \pi_j, \sum_{j=1}^m \pi_j = 1, \pi_j \geq 0, \quad j = 1, \dots, m \quad i = 1, \dots, n. \quad (1)$$

Where x_0^i represents the current (known) state of the world for the i^{th} asset, x_j^i represents the j^{th} realization of the i^{th} asset in the next time period (these are generated by the simulations mentioned above). r is the risk-less rate of return. The π_j are called the risk neutral probabilities. According to a fundamental result of Harisson and Kerps [17], the existence of the risk neutral probabilities is enough to guarantee that the scenario tree has the desired property. In the second step, we solve the least squares problem with some of the states fixed to the states calculated in the first step. In other words, we solve the following problem:

$$\begin{aligned} \min_{\omega, p} \sum_{i=1}^n \left(\sum_{j=1}^k p_j m_i(\omega_j) + \sum_{l=1}^m p_l m_i(\hat{\omega}_l) - \mu_i \right)^2 \\ \text{s.t.} \quad \sum_{j=1}^{k+m} p_j = 1 \quad p_j \geq 0 \quad j = 1, \dots, k+m \end{aligned} \quad (2)$$

In the problem above, $\hat{\omega}$ are fixed. Solving the preceding problem guarantees a scenario tree that is arbitrage free.

3 Portfolio Selection

In this section we describe the portfolio selection problem when higher order terms are taken into account. The classical mean-variance approach to portfolio analysis seeks to balance risk (measured by variance) and reward (measured by expected value). There are many ways to specify the single period problem. We will be using the following basic model:

$$\begin{aligned} & \min_w -\alpha\mathbb{E}[w] + \beta\mathbb{V}[w] \\ & \text{s.t. } \sum_{i=1}^n w_i = 1 \quad l_i \leq w_i \leq u_i \quad i = 1, \dots, n. \end{aligned} \tag{3}$$

Where $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ represent the mean rate of return and its variance respectively. The single constraint is known as the *budget constraint* and it specifies the initial wealth (without loss of generality we have assumed that this is one). The α and β are positive scalars, and are chosen so that $\alpha + \beta = 1$. They specify the DMs preferences, i.e. $\alpha = 1$ means that the DM is risk-seeking, while $\beta = 1$ implies that the DM is risk averse. Any other selection of the parameters will produce a point on the efficient frontier. The decision variable (w) represents the commitment of the DM to a particular asset. Note that this problem is a convex quadratic programming problem for which very efficient algorithms exists. The interested reader is referred to the review in [13] for more information regarding the Markowitz model.

We propose an extension of the mean-variance model using higher order moments. The vector optimization problem can be formulated as a standard non-convex optimization problem using two additional scalars to act as weights. These weights are used to enforce the DMs preferences. The problem is then formulated as follows:

$$\begin{aligned} & \min_w -\alpha\mathbb{E}[w] + \beta\mathbb{V}[w] - \gamma\mathbb{S}[w] + \delta\mathbb{K}[w] \\ & \text{s.t. } \sum_{i=1}^n w_i = 1 \quad l_i \leq w_i \leq u_i \quad i = 1, \dots, n. \end{aligned} \tag{4}$$

Where $\mathbb{S}[\cdot]$ and $\mathbb{K}[\cdot]$ represent the skewness and kurtosis of the rate of return respectively. γ and δ are positive scalars. The four scalar parameters are chosen so that they sum to one. Positive skewness is desirable (since it corresponds to higher returns albeit with low probability) while kurtosis is undesirable since it implies that the DM is exposed to more risk. The model in (4) can be extended to multiple periods while maintaining the same structure (non convex objective and linear constraints). The numerical solution of (2) and (4) will be discussed in the next two sections.

4 A Stochastic Optimization Algorithm

The models described in the previous section can be written as:

$$\begin{aligned} & \min_x f(x) \\ & \text{s.t. } Ax = b \\ & \quad x \geq 0. \end{aligned}$$

A well known method for obtaining a solution to an unconstrained optimization problem is to consider the following Ordinary Differential Equation (ODE):

$$dX(t) = -\nabla f(X(t))dt. \tag{5}$$

By studying the behavior of $X(t)$ for large t , it can be shown that $X(t)$ will eventually converge to a stationary point of the unconstrained problem. A review of, so called, continuous-path methods can be found in [18]. A deficiency of using (5) to solve optimization problems, is that it will get trapped in local minima. In order to allow the trajectory to escape from local minima, it has been proposed by various authors (e.g. [7, 8, 9, 10]) to add a stochastic term that would allow the trajectory to “climb” hills. One possible augmentation to (5) that would enable us to escape from local minima is to add noise. One then considers the *diffusion process*:

$$dX(t) = -\nabla f(X(t))dt + \sqrt{2T(t)}dB(t). \quad (6)$$

Where $B(t)$ is the standard Brownian motion in \mathbb{R}^n . It has been shown in [8, 9, 10], under appropriate conditions on f , and $T(t)$, that as $t \rightarrow \infty$, the transition probability of $X(t)$ converges (weakly) to a probability measure Π . The latter, has its support on the set of global minimizers.

For the sake of argument, suppose we did not have any linear constraints, but only positivity constraints. We could then consider enforcing the feasibility of the iterates by using a barrier function. According to the algorithmic framework sketched-out above, we could obtain a solution to our (simplified) problem, by following the trajectory of the following SDE:

$$dX(t) = -\nabla f(X(t))dt + \mu X(t)^{-1}dt + \sqrt{2T(t)}dB(t). \quad (7)$$

Where $\mu > 0$, is the barrier parameter. By X^{-1} , we will denote an n -dimensional vector whose i^{th} component is given by $1/X_i$. Having used a barrier function to deal with the positivity constraints, we can now introduce the linear constraints into our SDE. This process has been carried out in [6] using the projected SDE:

$$dX(t) = P[-\nabla f(X(t)) + \mu X(t)^{-1}]dt + \sqrt{2T(t)}PdB(t). \quad (8)$$

Where, $P = I - A^T(AA^T)^{-1}A$. The proposed algorithm works in a similar manner to gradient projection algorithms. The key difference is the addition of a barrier parameter for the positivity of the iterates, and a stochastic term that helps the algorithm escape from local minima.

The global optimization problem can be solved by fixing μ , and following the trajectory of (8) for a suitably defined function $T(t)$. After sufficiently enough time passes, we reduce μ , and repeat the process. The proof that following the trajectory of (8) will eventually lead us to the global minimum appears in [6]. Note that the projection matrix for the type of constraints we need to impose for our models is particularly simple. For a constraint of the type: $\sum_{i=1}^n x_i = 1$ the projection matrix is given by:

$$P_{ij} = \begin{cases} -\frac{1}{n} & \text{if } i \neq j, \\ \frac{n-1}{n} & \text{otherwise.} \end{cases}$$

5 Numerical Experiments

The algorithm described in the previous section was implemented in C++. Before we discuss our numerical results we provide some useful implementation details. From similar studies in the unconstrained case[7] and box constrained case[19], we know that a deficiency of stochastic methods (of the type proposed in this paper) is that they require a large number of function evaluations. The reason of this shortcoming is that the annealing schedule has to be sufficiently slow in order to allow the trajectory to escape from local minima. Therefore, whilst there are many sophisticated methods for the numerical solution of SDEs, we decided to use the cheaper stochastic Euler method. The latter method is a generalization of the well known Euler method, for ODEs, to the stochastic case. The main iteration is given by:

$$X(t + 1) = X(t) + P[-\nabla f(X(t)) + \mu X(t)^{-1}]\Delta t + \sqrt{2T(t)\Delta t}Pu.$$

Where Δt is the discretized step length parameter, u is a standard Gaussian vector, i.e. $u \sim N(0, I)$, and $X(0)$ is chosen to be strictly feasible.

The algorithm starts by dividing the discretized time into k periods. Following a single trajectory will be too inefficient. Therefore, starting from a single strictly feasible point the algorithm generates m different trajectories. After a single period elapses, we remove the worst performing trajectory. Since, all trajectories generate feasible points, we can assess the quality of the trajectory by the best objective function value achieved on the trajectory. We then randomly select one of the remaining trajectories, and duplicate it. At this stage we reduce the noise coefficient of the duplicated trajectory.

When all the periods have been completed, in the manner described above, we count this event as one iteration. At this point we reduce the barrier parameter. This parameter is started at $\mu = 0.1$, and reduced by 0.75 at every iteration. We then repeat the same process, with all the trajectories starting from the best point found so far. If the current incumbent solution vector remained the same for more than l iterations ($l > 4$, in our implementation) then we reset the noise to its initial value. The algorithm terminates when the noise term is smaller than a predefined value ($0.1e - 4$) or when after five successive resets of the noise term, no improvement could be made. In our implementation we used two trajectories, two periods (each of length $20e4$). We used an initial value of 10 for the annealing schedule, and reduced it by 0.6 at every iteration. The same parameters were used for all the simulations.

In table 1 we show the computational effort required to compute the sufficient set of states required to guarantee the arbitrage free property of the scenario tree. The numbers shown are averages of 50 runs. It is clear from table 1 that it is relatively easy to find the sufficient states. However, as the number of assets increases the number of states also increases. While in a single period model this does not cause much computational burden, it suggests that it will lead to a state explosion in the multi-period case. In the future we plan to investigate the approach of finding the state that causes the arbitrage opportunity and eliminating/modifying it rather than just adding more states.

Table 1. States added to guarantee arbitrage free tree

Assets	States Added	Time (secs)
2	4	0
5	14	0
10	27	0.01
15	77	0.14
20	170	0.79

Table 2. Solution Times Moment Problem

Assets	Time (secs)	Variables
2	820	105
5	1230	111
10	4160	124
15	23316	184
20	52544	242

Table 3. Solution Times Portfolio Selection

Assets	M-V	M-V-S	M-V-K	M-V-S-K
2	0.01	56	36	34
5	0.3	138	101	204
10	1.3	250	195	195
15	4	375	433	519
20	14.8	551	776	762

In table 2 we show the time required to solve the moment problem. The number of variables differ from one run to the next. This is because the number of states that are needed to guarantee the arbitrage free property differ from run to run (since they are randomly generated). In all runs we added fifty more (non-constant) states. The resulting problem given by (2) was then solved using the algorithm described above. The times shown are the averages for ten runs for problems with 2, 5, and 10 assets. Due to the large amount of time required to solve the larger problems (15 and 20 assets) the times reported are from a single run. Table 3 details the time required to generate a point on the efficient frontier for the four versions of the portfolio selection problem we considered in this paper. The first is the mean-variance (M-V) model, this is obtained by setting $\gamma = \delta = 0$ (we used this model to test the quality of the solutions provided by the algorithm). Similarly M-V-S, M-V-K and M-V-S-K stand for Mean-Variance-Skewness, Mean-Variance-Kurtosis and Mean-Variance-Skewness-Kurtosis respectively. We realize that providing the computation times is not the best way to judge the speed of an algorithm. However, one of the aims of this paper is to show how one can use a stochastic global optimization algorithm to solve a financial planning problem. Even though there are many open questions, and some of our assumptions may be too stringent, we believe that the computation times tabulated below show the feasibility of this approach. In figures 1 and 2 we show some 3-dimensional efficient frontiers for the M-V-S and M-V-K problems respectively. The gaps that appear in the

frontiers are due to the way we generate the frontier. If we used constraints, instead of weights, to express the preferences of the DM, then we believe that the frontier would look more smooth. However, a formulation using constraints would lead to an optimization problem that could not be solved by our global optimization solver. We plan to address this deficiency in the future. Figures 3 and 4 show efficient frontiers using the M-V-S-K model. In figure 3 we plot the first three measures of interest, while in figure 4 we plot the mean, variance and kurtosis of the portfolio.

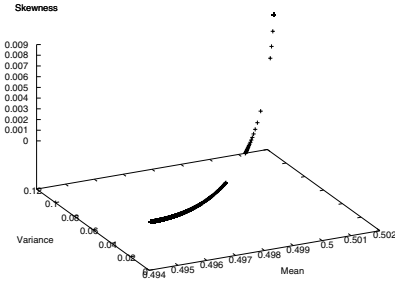


Fig. 1. Mean-Variance-Skewness

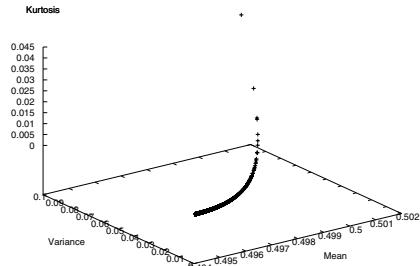


Fig. 2. Mean-Variance-Kurtosis

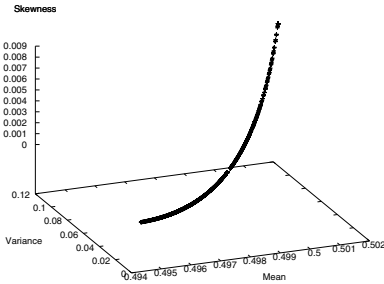


Fig. 3. Mean-Variance-Skewness-Kurtosis

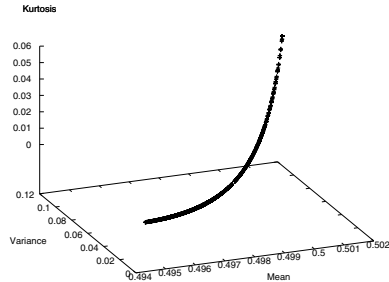


Fig. 4. Mean-Variance-(Skewness)-Kurtosis

6 Conclusions

We considered the computational challenges associated with the global optimization of a financial planning model. We proposed a simple extension to the classical Markowitz model; in the proposed model higher order moments were included using scalar weights. The scenario generation problem was also addressed by matching the first four central moments of the postulated distribution. We also addressed the issue of imposing the arbitrage free property to the generated scenario tree. A stochastic algorithm was proposed for the two models. Our initial numerical results show that problems of realistic size can be solved using the proposed framework.

References

1. Markowitz, H.M.: Portfolio selection. *J. Finance* **7** (1952) 77–91
2. Markowitz, H.M.: The utility of wealth. *J. Polit. Econom.* (1952) 151–158
3. Dupacova, J., Consigli, G., Wallace, S.: Scenarios for multistage stochastic programs. *Ann. Oper. Res.* **100** (2000) 25–53 (2001)
4. Gülpınar, N., Rustem, B., Settergren, R.: Simulation and optimization approaches to scenario tree generation. *J. Econom. Dynam. Control* **28** (2004) 1291–1315
5. Prékopa, A.: Stochastic programming. Volume 324 of Mathematics and its Applications. Kluwer Academic Publishers Group, Dordrecht (1995)
6. Parpas, P., Rustem, B., Pistikopoulos, E.N.: Linearly constrained global optimization and stochastic differential equations. Accepted in the *J. Global Optim.* (2006)
7. Aluffi-Pentini, F., Parisi, V., Zirilli, F.: Global optimization and stochastic differential equations. *J. Optim. Theory Appl.* **47** (1985) 1–16
8. Chiang, T., Hwang, C., Sheu, S.: Diffusion for global optimization in \mathbf{R}^n . *SIAM J. Control Optim.* **25** (1987) 737–753
9. Geman, S., Hwang, C.: Diffusions for global optimization. *SIAM J. Control Optim.* **24** (1986) 1031–1043
10. Gidas, B.: The Langevin equation as a global minimization algorithm. In: *Disordered systems and biological organization* (Les Houches, 1985). Volume 20 of NATO Adv. Sci. Inst. Ser. F Comput. Systems Sci. Springer, Berlin (1986) 321–326
11. Konno, H.: Applications of global optimization to portfolio analysis. In C., A., P., H., G., S., eds.: *Essays and Surveys in Global Optimization*. Springer (2005) 195–210
12. Maranas, C.D., Androulakis, I.P., Floudas, C.A., Berger, A.J., Mulvey, J.M.: Solving long-term financial planning problems via global optimization. *J. Econom. Dynam. Control* **21** (1997) 1405–1425 Computational financial modeling.
13. Steinbach, M.C.: Markowitz revisited: mean-variance models in financial portfolio analysis. *SIAM Rev.* **43** (2001) 31–85 (electronic)
14. Parpas, P., Rustem, B.: Entropic regularization of the moment problem. Submitted (2005)
15. Bertsimas, D., Sethuraman, J.: Moment problems and semidefinite optimization. In: *Handbook of semidefinite programming*. Volume 27 of Internat. Ser. Oper. Res. Management Sci. Kluwer Acad. Publ., Boston, MA (2000) 469–509
16. Klaassen, P.: Discretized reality and spurious profits in stochastic programming models for asset/liability management. *E.J of Op. Res.* **101** (1997) 374–392
17. Harrison, J., D.M., K.: Martingales and arbitrage in multiperiod securities markets. *J. Econom. Theory* **20** (1979) 381–408
18. Zirilli, F.: The use of ordinary differential equations in the solution of nonlinear systems of equations. In: *Nonlinear optimization, 1981* (Cambridge, 1981). NATO Conf. Ser. II: Systems Sci. Academic Press, London (1982) 39–46
19. Recchioni, M.C., Scoccia, A.: A stochastic algorithm for constrained global optimization. *J. Global Optim.* **16** (2000) 257–270

A Generalized Fuzzy Optimization Framework for R&D Project Selection Using Real Options Valuation*

E. Ertugrul Karsak

Industrial Engineering Department, Galatasaray University,
Ortaköy, Istanbul 80840, Turkey
ekarsak@gsu.edu.tr

Abstract. Global marketplace and intense competition in the business environment lead organizations to focus on selecting the best R&D project portfolio among available projects using their scarce resources in the most effective manner. This happens to be a *sine qua non* for high technology firms to sharpen their competitive advantage and realize long-term survival with sustainable growth. To accomplish that, firms should take into account both the uncertainty inherent in R&D using appropriate valuation techniques accounting for flexibility in making investment decisions and all possible interactions between the candidate projects within an optimization framework. This paper provides a fuzzy optimization model for dealing with the complexities and uncertainties regarding the construction of an R&D project portfolio. Real options analysis, which accounts for managerial flexibility, is employed to correct the deficiency of traditional discounted cash flow valuation that excludes any form of flexibility. An example is provided to illustrate the proposed decision approach.

1 Introduction

In this paper, the research and development (R&D) project selection problem is addressed. R&D project selection examines the allocation of company's scarce resources such as budget, manpower, etc. to a set of proposals to enhance its strategic performance on a scientific and technological basis. R&D is crucial for a company's competitive advantage, survival and sustainable growth. R&D enables the company to develop new products or services, enhance existing ones, and increase efficiency while lowering cost of the production processes. This paper focuses on the problem of selecting a portfolio of R&D projects when both vagueness and uncertainty in data and interactions between candidate projects exist.

For the case of crisp data, early work dates back to Weingartner [16], and since then R&D project selection has been an active area of research for academics and practitioners. Liberatore [10] proposed a decision framework based on the analytic hierarchy process (AHP) and integer programming for R&D project selection. Inadequate representation of project interdependencies, and the inability to incorporate the uncertainty inherent in projects and interactions between projects are the major

* This research has been financially supported by Galatasaray University Research Fund.

shortcomings of the previously proposed analytical models for R&D project selection. Within the last decade, researchers, in particular, addressed the proper treatment of project interdependencies. Schmidt [13] presented a model that considered benefit, outcome and resource interactions and proposed a branch and bound algorithm to obtain a solution for the nonlinear integer programming problem with quadratic constraints. Meade and Presley [12] proposed the utilization of the analytic network process (ANP) that enables the decision-maker to take into consideration interdependencies among criteria and candidate projects. One should note the resource feasibility problem that may be encountered while using the ANP by itself. Lee and Kim [8] presented an integrated application of the ANP and zero-one goal programming for information system project selection to consider resource feasibility as well as project interdependence.

Although the aforementioned valuable contributions considered interactions between projects, they all relied on crisp data. R&D projects comprise a high degree of uncertainty, which generally precludes the availability of obtaining exact data regarding benefit, resource usage, and interactions between projects. Fuzzy set theory appears as a useful tool to account for vagueness and uncertainty inherent in the R&D project selection process. Recently, a model that handles fuzzy benefit and resource usage assuming that a project can influence at most one other was developed; however, an optimization procedure was not presented to solve the proposed model [7].

Furthermore, the research studies cited above use the traditional discounted cash flow (DCF) techniques such as net present value (NPV) in its static form for calculating the benefits from R&D investments. Lately, options valuation approach has been proposed as a more suitable alternative for determining the benefits from R&D projects [9]. Options approach deviates from the conventional DCF approach in that it views future investment opportunities as rights without obligations to take some action in the future [3]. The asymmetry in the options expands the NPV to include a premium beyond the static NPV calculation, and thus presumably increase the total value of the project and the probability of justification. Carlsson and Fullér [1] further extended the use of options approach in R&D project valuation by considering the possibilistic mean and variance of fuzzy cash flow estimates.

This paper presents a novel fuzzy formulation for R&D project selection accounting for project interactions with the objective of maximizing the net benefit based on expanded net present value which incorporates the real options inherent in R&D projects in a fuzzy setting. Although a fuzzy optimization model is provided for R&D portfolio selection in [14], the project interactions are completely ignored. Compared with the real options valuation procedures delineated in [1, 14], the valuation approach utilized in this paper models exercise price as a stochastic variable enabling to deal with technological uncertainties in real options analysis, and considers both the benefits of keeping the development option alive and the opportunity cost of delaying development. Moreover, this paper's focus is not limited to valuation of R&D projects using fuzzy cash flow estimates since the proposed optimization framework enables constructing an optimal portfolio of R&D projects considering the commonly encountered project interdependencies regarding resource usage. The proposed model will lead to a binary integer program with nonlinear constraints. In this paper, a solution procedure based on linearization of the nonlinear constraints is provided and thus the resulting linear problem can be solved with widely available solvers. The

proposed optimization approach is also advantageous compared with heuristics in that the obtained solution is the global optimum of the problem.

The rest of the paper is organized as follows. Section 2 reviews the sequential exchange options for valuing R&D projects. Section 3 outlines the approach to incorporate fuzzy cash flows into the valuation methodology. A fuzzy optimization model with nonlinear constraints which is later converted into a crisp linear binary integer program is introduced in Section 4. A comprehensive example is presented in the subsequent section to illustrate the application of the proposed framework. Finally, conclusions and directions for future research are provided in Section 6.

2 Real Options Approach to Valuation of R&D Projects

An option is the right, but not the obligation, to buy (if a call) or sell (if a put) a particular asset at a specified price on or before a certain expiration date. The buyer of an option may choose to exercise his right and take a position in the underlying asset while the option seller, also known as the option writer, is contractually obligated to take the opposite position in the underlying asset if the buyer exercises his right. The price at which the buyer of an option may buy or sell the underlying asset is the exercise price. An American option can be exercised at any time prior to expiration, while a European option allows exercise only on its expiration date. An American exchange option, which can be cited among options with more complicated payoffs than the standard European or American calls and puts, gives its owner the right to exchange one asset for another at any time up to and including expiration.

While financial options are options on financial assets, real options are opportunities on real assets that can provide management with valuable operating flexibility and strategic adaptability. Akin to financial options, real options enable their owners to revise future investment and operating decisions according to the market conditions. A substantial part of the market value of companies operating in volatile and unpredictable industries such as electronics, telecommunications, and biotechnology can be attributed to the real options that they possess [3]. Real options preclude the traditional passive analysis of investments, and imply active management approach with an ability to respond to changing conditions. Real options approach enables the firm to evaluate the project in a multi-stage context, providing the means to revise the decisions based on new information.

It is reported that American sequential exchange options provide a more realistic valuation of R&D projects compared with other option models when R&D projects incorporate stages of research and/or sequential investment opportunities [9]. In this paper, an efficient method for valuing American sequential exchange options when both underlying assets pay dividends continuously and there exists a possibility of early exercise is employed. The earlier work on exchange options dates back to Margrabe [11], who derived a pricing equation for the exchange options on non-dividend-paying assets. Although elegant by its ability to model exercise price as a stochastic variable enabling to deal with technological uncertainties in real options analysis, easy to use formula developed by Margrabe [11] falls short of incorporating dividends into the analysis, which may especially be crucial in valuation of real options due to the fact that the underlying assets are generally not traded.

Here, an option to exchange asset D for asset V at time T is considered. Asset D is referred as the delivery asset, and asset V as the optioned asset. The payoff to this European option at time T is given as $\max(0, V_T - D_T)$, where V_T and D_T are the underlying assets' terminal prices. The asset prices prior to expiration, i.e. V and D , are assumed to follow geometric Brownian motion as

$$\begin{aligned} \frac{dV}{V} &= (\alpha_v - \delta_v)dt + \sigma_v dZ_v, \\ \frac{dD}{D} &= (\alpha_d - \delta_d)dt + \sigma_d dZ_d, \\ \text{cov}\left(\frac{dV}{V}, \frac{dD}{D}\right) &= \rho_{vd} \sigma_v \sigma_d dt. \end{aligned} \tag{1}$$

where α_v and α_d are the expected rates of return on the two assets, δ_v and δ_d are the corresponding dividend yields, σ_v^2 and σ_d^2 are the respective variance rates, and dZ_v and dZ_d are increments of the Wiener processes at time t ($t \in [0, T]$). The rates of price change, i.e. $\left(\frac{dV}{V}\right)$ and $\left(\frac{dD}{D}\right)$, can be correlated with the correlation coefficient denoted by ρ_{vd} . The parameters δ_v , δ_d , σ_v , σ_d , and ρ_{vd} are non-negative constants. δ , which denotes the difference in dividend yields, can be defined as $\delta_v - \delta_d$.

In this paper, the model proposed by Carr [2] for valuing American exchange options on dividend-paying assets is used. Carr [2] generalized the solution of Geske and Johnson [4], which was initially developed for valuing an American put option, to American exchange options on assets with continuous dividends. Geske and Johnson [4] viewed an American put option as the limit to a sequence of pseudo-American puts. A pseudo-American option can only be exercised at a finite number of discrete exercise points. In the limiting case, the value of a pseudo-American option approaches the exact value of a true American put option. Geske and Johnson [4] achieved accuracy by considering put options which can be exercised at a small number of discrete time points, and then employed the values obtained at these exercise dates to extrapolate to the value of a put option that can be exercised at any date. The details of the valuation formula for the general pseudo-American exchange option are not provided here due to limited space. The reader may refer to Karsak and Özogul [6] for a detailed presentation.

3 Using Fuzzy Sets for Modeling Uncertainty in Project Selection

The fuzzy set theory deals with problems in which a source of imprecision and vagueness is involved. A fuzzy set can be defined mathematically by assigning to each possible individual in the universe of discourse a value representing its grade of membership in the fuzzy set. This grade corresponds to the degree to which individual is compatible with the concept represented by the fuzzy set. A convex and normalized

fuzzy set defined on \mathfrak{X} with a piecewise continuous membership function is called a fuzzy number. Uncertainty and imprecision in parameters such as cash flow estimates can be incorporated into the R&D project selection framework using fuzzy numbers. $\tilde{A} = (c_L, c_R, s_L, s_R)$ is a trapezoidal fuzzy number with the membership function defined as

$$f_{\tilde{A}}(x) = \begin{cases} (x - (c_L - s_L)) / s_L, & c_L - s_L \leq x \leq c_L \\ 1, & c_L \leq x \leq c_R \\ ((c_R + s_R) - x) / s_R, & c_R \leq x \leq c_R + s_R \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where c_L and c_R are the left and right core values, and s_L and s_R are the left and right spreads, respectively. The support of \tilde{A} is $(c_L - s_L, c_R + s_R)$. If $c_L = c_R = c$, the resulting fuzzy number is a triangular fuzzy number denoted as $\tilde{A} = (c, s_L, s_R)$. Further, when $s_L = s_R = s$, a symmetric triangular fuzzy number $\tilde{A} = (c, s)$ is obtained.

The possibilistic mean and variance of a trapezoidal fuzzy number \tilde{A} are defined as [1]

$$E(\tilde{A}) = \frac{c_L + c_R}{2} + \frac{s_R - s_L}{6}, \tag{3}$$

$$Var(\tilde{A}) = \frac{(c_R - c_L)^2}{4} + \frac{(c_R - c_L)(s_L + s_R)}{6} + \frac{(s_L + s_R)^2}{24}. \tag{4}$$

4 Fuzzy Optimization Framework for R&D Project Selection

This paper considers constructing an R&D project portfolio, where there are m candidate R&D projects. The binary decision variable x_i ($i = 1, \dots, m$) corresponds to the i th R&D project, where $x_i = 1$ if R&D project i is selected and $x_i = 0$ otherwise. The objective is to maximize the total net benefit obtained from the R&D project portfolio. Resource constraints related to the initial expenditures for the R&D projects and the skilled workforce (in man-hours) are considered as well as the interdependencies among the R&D projects regarding the use of these resources. There is an estimate for budget limit for initial expenditures (\tilde{T}_B) and an estimate for skilled workforce limit required for the development phase of R&D projects (\tilde{T}_W). Both of these estimates as well as the estimates for resource usages and shared resources for the R&D projects are represented as fuzzy numbers due to the imprecise nature of the problem. The proposed model also enables to account for project contingencies, which indicate a project cannot be implemented unless a related project is also selected. Other restrictions regarding the construction of the R&D project portfolio such as mutually exclusive projects or mandated projects can be readily appended to the proposed model.

$$\max Z^* = \sum_{i=1}^m V_i^e x_i, \tag{5}$$

subject to

$$\sum_{i=1}^m \tilde{C}_i x_i - \sum_{i=1}^{m-1} \sum_{j=i+1}^m \tilde{C}_{ij} x_i x_j + \sum_{i=1}^{m-2} \sum_{j=i+1}^{m-1} \sum_{k=j+1}^m \tilde{C}_{ijk} x_i x_j x_k \leq \tilde{T}_B, \tag{6}$$

$$\sum_{i=1}^m \tilde{W}_i x_i - \sum_{i=1}^{m-1} \sum_{j=i+1}^m \tilde{W}_{ij} x_i x_j + \sum_{i=1}^{m-2} \sum_{j=i+1}^{m-1} \sum_{k=j+1}^m \tilde{W}_{ijk} x_i x_j x_k \leq \tilde{T}_W, \tag{7}$$

$$\sum_{i \in Y_j} x_i \geq |Y_j| x_j, \quad j \in \Theta_Y, \tag{8}$$

$$x_i \in \{0,1\}, \quad i = 1, \dots, m. \tag{9}$$

Formula (5) represents the objective of maximizing the total net benefit of the R&D project portfolio, where V_i^e denotes the net benefit obtained from project i using the expanded net present value (ENPV). Both formulae (6) and (7) represent resource constraints, which enable project interactions to be taken into account. Uncertain initial expenditure and shared initial expenditure parameters are denoted respectively as \tilde{C}_i and $\tilde{C}_{ij}, \tilde{C}_{ijk}$, while \tilde{W}_i and $\tilde{W}_{ij}, \tilde{W}_{ijk}$ represent fuzzy workforce and fuzzy shared workforce parameters, respectively, for the R&D projects. Although the current formulation assumes that interactions exist among at most three R&D projects, it can be easily extended to include higher number of interdependent projects. In addition to the resource constraints, the formulation includes contingency constraints given by formula (8) indicating that the implementation of the project j is contingent upon the implementation of all the projects in $Y_j \subset \{1, \dots, m\}$, where $|Y_j|$ indicates the cardinal of Y_j and $\Theta_Y \subset \{1, \dots, m\}$.

The formulation given above is a fuzzy nonlinear integer programming model and can be linearized using the approach delineated in [15]. For instance, the nonlinear term $x_i x_j$ can be linearized by introducing a new variable $x_{ij} := x_i x_j$, where $x_{ij} \in \{0,1\}$, and appending the following linear constraints to the model:

$$x_i + x_j - x_{ij} \leq 1, \tag{10}$$

$$-x_i - x_j + 2x_{ij} \leq 0. \tag{11}$$

After performing the linearization, the fuzzy linear integer programming formulation can be converted to a crisp mathematical programming model using the possibility theory. Formulae (6) and (7) that incorporate fuzzy parameters can be

rewritten as crisp constraints employing the possibilistic approach [5]. For example, an inequality constraint given as

$$\tilde{a}_{1j}x_1 + \tilde{a}_{2j}x_2 + \dots + \tilde{a}_{mj}x_m \leq \tilde{b}_j \tag{12}$$

can be rewritten as

$$\sum_{i=1}^m a_{ij}^{cR}x_i + \lambda_j \sum_{i=1}^m a_{ij}^{sR}x_i \leq b_j^{cR} + (1 - \lambda_j)b_j^{sR}, \tag{13}$$

where λ_j is the satisfaction degree of the constraint, a_{ij}^{cR}, b_j^{cR} denote the right core values of $\tilde{a}_{ij}, \tilde{b}_j$, and a_{ij}^{sR}, b_j^{sR} denote their right spreads, respectively.

5 Illustrative Example

In this section, we consider a technology firm analyzing six R&D project alternatives, where each R&D project consists of a two-stage investment, namely the initial stage and the development stage. The company acquires the right to make a development investment by making the initial investment for each R&D project. Due to imprecise and uncertain nature of R&D investments, project cash flow estimates regarding the development stage are given as fuzzy numbers in Table 1. Although the methodology delineated in the previous section enables to use trapezoidal fuzzy numbers, in order to save space, fuzzy cash flow estimates are provided as symmetric triangular fuzzy numbers $\tilde{A}_j = (c, s)$ where c is the most likely (core) value and s is the spread, respectively. Resource data for the R&D project alternatives and data related to the project interactions are provided in Table 2 and Table 3, respectively. Project 2 is assumed to be contingent upon the implementation of project 5.

Table 1. Fuzzy cash flow estimates (in thousands of dollars) regarding the development stage of the R&D project alternatives

	Project 1	Project 2	Project 3	Project 4	Project 5	Project 6
A ₀	(-27000,6000)	(-30000,6500)	(-35000,7500)	(-40000,8500)	(-25000,5500)	(-41000,9000)
A ₁	(7000,1500)	(8000, 1800)	(11500,2700)	(7500,1800)	(4000,900)	(11000,2500)
A ₂	(8000,1800)	(9000, 2100)	(14000,3200)	(9300,2400)	(5400,1300)	(13000,3100)
A ₃	(8500,2200)	(9500, 2300)	(12000,2900)	(10500,2500)	(6500,1600)	(13500,3400)
A ₄	(8000,2100)	(10500,2400)	(10600,2800)	(11500,3000)	(7000,1700)	(16500,4000)
A ₅	(6700,1700)	(8800, 2500)	(9500,2500)	(14000,3200)	(7500,1900)	(17500,4300)
A ₆	(6000,1600)	(8000, 2200)	(8500, 2200)	(12500,3000)	(9000,2200)	(15000,3800)

Equation (3) is employed to compute the expected value of the development stage investment and the expected value of returns from the development stage investment for each R&D project alternative. The standard deviation of the rate of change of the

returns from the development stage investment and the standard deviation of the rate of change of the development stage investment are taken to be 0.1 and 0.3, respectively. It is assumed that the correlation between development investment and expected value of returns from development investment is 0.5 for each R&D project. The time interval in which the development investment can be realized is taken to be two years. The opportunity cost of delaying development investment (δ_v) is set as a constant proportional to expected value of returns from the development investment as 0.04 while the depreciation of development investment is determined as a constant proportional to value of development investment as $\delta_d = 0.02$. In practice, δ_v depends on competitive intensity and market structure characteristics in addition to the anticipated increase in demand and can be measured from market information using econometric methods, whereas δ_d can be estimated based on expert opinion [6].

Table 2. Resource data for the R&D project alternatives

Projects	Initial expenditures (\$)	Workforce (in man-hours)
1	6,000,000	(200,000, 20,000)
2	9,500,000	(240,000, 20,000)
3	13,500,000	(300,000, 30,000)
4	7,000,000	(320,000, 30,000)
5	3,000,000	(160,000, 20,000)
6	20,000,000	(360,000, 40,000)

Table 3. Data regarding the R&D project interdependencies

Interdependent projects	Shared initial expenditures (in thousands of dollars)	Shared workforce (in man-hours)
1, 3	(2,000, 400)	(30,000, 6,000)
1, 6	(1,500, 200)	(40,000, 8,000)
2, 4	(2,500, 400)	(50,000, 10,000)
2, 5	(2,000, 200)	(24,000, 4,000)
3, 6	(3,000, 500)	(40,000, 10,000)
2, 4, 5	(2,000, 300)	(30,000, 8,000)

NPV for the development stage of each R&D project is calculated using an interest rate of 10%. ENPV for the development stage of each R&D project is obtained employing the real options valuation approach delineated in Section 2. The difference between ENPV and NPV gives the option value for each R&D project. The net benefit obtained from each R&D project using ENPV is computed using equation (14).

$$\text{Net Benefit (R\&D project)} = \text{Initial Expenditure} + \text{ENPV (development stage)} \quad (14)$$

The results of these valuations are reported in Table 4. In Table 4, C denotes the initial expenditure, V represents the net benefit calculation based on NPV, and V^e

denotes the net benefit using ENPV for the respective R&D project. As shown in Table 4, the net benefit calculation based on NPV results in negative figures for projects 1, 2, 4 and 5, whereas the net benefit using ENPV yields positive results for every R&D project alternative. In other words, an optimization model that maximizes the total net benefit based on NPV would eliminate projects 1, 2, 4 and 5 as a result of an erroneous valuation procedure ignoring any form of flexibility.

Table 4. Valuation results (in dollars) for the R&D project alternatives

	Project 1	Project 2	Project 3	Project 4	Project 5	Project 6
NPV	5,372,507	8,999,775	13,977,295	5,996,415	2,500,989	20,489,506
ENPV	6,821,550	9,900,180	14,527,765	8,695,080	4,614,975	20,748,132
Option Value	1,449,043	900,405	550,470	2,698,665	2,113,986	258,626
$V = NPV - C$	-627,493	-500,225	477,295	-1,003,585	-499,011	489,506
$V^e = ENPV - C$	821,550	400,180	1,027,765	1,695,080	1,614,975	748,132

Considering $\tilde{T}_B = (30,000,000, 4,000,000)$ dollars, $\tilde{T}_W = (1,000,000, 100,000)$ hours and a satisfaction degree of 0.8 for the resource constraints ($\lambda_j = 0.8$), the optimal solution of the crisp mathematical programming model, which is obtained by applying the linearization scheme delineated in Section 4 and the possibility theory to the fuzzy nonlinear integer programming formulation represented by formulae (5)-(9), is determined as projects 1, 2, 4 and 5. The optimal R&D project portfolios for varying satisfaction degrees of resource constraints are presented in Table 5. As can be seen in Table 5, the portfolio including projects “1, 3, 4, 5”, which yields a higher net benefit figure compared with the portfolio consisting of projects “1, 2, 4, 5”, becomes infeasible as the satisfaction degree for resource constraints increases to 0.8. It is also worth noting that both the real options valuation approach and the optimization framework used in this paper enable further sensitivity analyses regarding parameter flexibilities.

Table 5. Optimal solutions for varying λ_j values

λ_j	Z^*	Selected projects
0.6	5,159,370	1, 3, 4, 5
0.7	5,159,370	1, 3, 4, 5
0.8	4,531,785	1, 2, 4, 5
0.9	4,531,785	1, 2, 4, 5

6 Conclusions

This paper aims to develop a fuzzy optimization approach to select an R&D project portfolio while accounting for project interactions and determining the benefits resulting from the R&D investments using real options valuation in a fuzzy setting. Although a fuzzy approach to R&D project selection enables to hedge against the R&D

uncertainty, another important issue that needs to be considered is that traditional DCF valuation methods oftentimes undervalue the risky project. In this paper, American sequential exchange options are employed to address this problem. Since R&D projects generally involve phased research and sequential investment opportunities with uncertain expenditures as well as returns, the sequential exchange option model appears to be more suitable than other option pricing techniques. Sensitivity analysis with respect to parameters of the model, which is confined to a minimum here due to limited space, can be easily extended. A more efficient linearization scheme may also be applicable for the cases which include higher number of interdependent projects. The implementation of the proposed approach using real data remains as a future research objective.

References

1. Carlsson, C., Fullér, R.: A fuzzy approach to real option valuation. *Fuzzy Sets and Systems* 139 (2003) 297-312
2. Carr, P.: The valuation of sequential exchange opportunities. *The Journal of Finance* 43 (1988) 1235-1256
3. Dixit, A.K., Pindyck, R.S.: The options approach to capital investment. *Harvard Business Review* May-June (1995) 105-115
4. Geske, R., Johnson, H.E.: The American put option valued analytically. *The Journal of Finance* 39 (1984) 1511-1524
5. Inuiguchi, M., Ramik, J.: Possibilistic linear programming: a brief review of fuzzy mathematical programming and a comparison with stochastic programming in portfolio selection problem. *Fuzzy Sets and Systems* 111 (2000) 3-28
6. Karsak, E.E., Özogul, C.O.: An options approach to valuing expansion flexibility in flexible manufacturing system investments. *The Engineering Economist* 47 (2002) 169-193
7. Kuchta, D.: A fuzzy model for R&D project selection with benefit, outcome and resource interactions. *The Engineering Economist* 46 (2001) 164-180
8. Lee, J.W., Kim, S.H.: Using analytic network process and goal programming for interdependent information system project selection. *Computers & Operations Research* 27 (2000) 367-382
9. Lee, J., Paxson, D.A.: Valuation of R&D real American sequential exchange options. *R&D Management* 31 (2001) 191-201
10. Liberatore, M.: An extension of the analytic hierarchy process for industrial R&D project selection and resource allocation. *IEEE Trans. Eng. Manage.* 34 (1987) 12-18
11. Margrabe, W.: The value of an option to exchange one asset for another. *The Journal of Finance* 33 (1978) 177-186
12. Meade, L.M., Presley, A.: R&D project selection using the analytic network process. *IEEE Trans. Eng. Manage.* 49 (2002) 59-66
13. Schmidt, R.L.: A model for R&D project selection with combined benefit, outcome and resource interactions. *IEEE Trans. Eng. Manage.* 40 (1993) 403-410
14. Wang, J., Hwang, W.-L.: A fuzzy set approach for R&D portfolio selection using a real options valuation model. To appear in *Omega* (2006)
15. Watters, L.J.: Reduction of integer polynomial programming problems to zero-one linear programming problems. *Operations Research* 15 (1967) 1171-1174
16. Weingartner, H.M.: Capital budgeting of interrelated projects: survey and synthesis. *Management Science* 12 (1966) 485-516

Supply Chain Network Design and Transshipment Hub Location for Third Party Logistics Providers

Seungwoo Kwon¹, Kyungdo Park², Chulung Lee^{3,*}, Sung-Shick Kim³,
Hak-Jin Kim⁴, and Zhong Liang⁵

¹ Korea University Business School, Korea University

² College of Business Administration, Sogang University

³ Department of Industrial Systems and Information Engineering, Korea University
leecu@korea.ac.kr

⁴ Yonsei School of Business, Yonsei University

⁵ The Logistics Institute Asia Pacific, National University of Singapore

Abstract. Transshipment hubs are the places where cargo can be re-consolidated and transportation modes can be changed. Transshipment hubs are widely used in logistics industry. In this article, we consider a supply chain network design problem for a third party logistics provider in which we obtain the locations of transshipment hubs among promising hub locations for the objective of minimizing the total system cost that includes the transportation cost, the fixed cost and the processing cost. In the problem, the unit cost of transporting cargo between a pair of origin-destination is non-linear and is a decreasing step function, and each unit cargo through the transshipment hub is charged a fixed processing cost. The problem is considered a mixed integer programming problem. However, due to the complexity of the problem, a heuristic solution approach is proposed. The proposed solution is then implemented to a third party logistics provider and their experience shows that significant cost savings are obtained, compared with the current practice.

1 Introduction

Logistics industry has been a fast growing industry in the past forty years. The cost of the business logistics system globally is estimated to exceed US\$2 trillion in 1999 (The Logistics Institute Asia Pacific, 2004). In the United States, the logistics industry was worth more than \$920 billion, equivalent of approximately 10 percent of the gross domestic product, while the figures were 16 percent in 1980's. Logistics costs comprise more than 10-15% of the final product cost of finished goods. The total logistics cost included \$377 billion of the inventory carrying cost and \$585 billion of the transportation cost. From 1980 to 2000, with 1980 serving as the base, the inventory carrying cost has been declined by more than 50 percent, the transportation costs have been reduced by 22 percent and total logistics costs have decreased by 37 percent. The trends of more efficient inventory investment and fast cycle procurement have driven the productivity during the 1990s. Manufacturing and distribution

* Corresponding Author.

industries have realized that the logistics activity is not their core business and in order to achieve competitive edges, they must turn to supply chain management specialists whose core competencies are logistics and supply chain management. Regardless, for 1999, only less than 5% of the logistics activities were outsourced worldwide (Viswanadam et al., 2003). The outsourced segment of logistics expenditures by manufacturing and distributing businesses was estimated to be \$6 billion and increasing very fast. The centre in the growth is the third party logistics services providers. Third-party logistics service providers (3PLs) take over most of routine material handling activities such as warehousing, order picking, assembly, packaging and shipping, as well as the handling of returned parts and goods (Viswanadham et al., 2003). In addition, Third-party logistics providers are extending their business areas to marketing and customer relationship (especially in industries such as mobile phones and other high tech gadgets), assembly and ad-hoc manufacturing (for delayed differentiation), and supply hubs (supplier management). These new activities help them to create greater values (rather than merely cutting down operating costs) for their clients. Third-party logistics services grew by 24 percent in 2000. Dedicated contract carriage grew by 21 percent. Domestic transportation management grew by 21 percent in the face of competition from the so-called dot com freight exchanges. Warehouse based integrated services grew by 23 percent. Total contract logistics revenues grew by 24 percent or \$56.4 billion in 2000 (The Logistics Institute Asia Pacific, 2004).

In this study, we study a supply chain network design and control problem for a real-life third party logistics provider. Our objective is to obtain a high quality solution for their transshipment hub locations and transportation, simultaneously.

The rest of the paper proceeds as follow. Chapter 2 provides a literature review on the transshipment problem. Chapter 3 briefly discusses negotiation processes for the supply chain design. In Chapter 4, we discuss a quantitative model and the corresponding heuristic solution approach. The implementation issue and results are reported in Chapter 5.

2 Literature Review

Diks and De Kok (1996) investigate a two-echelon inventory system with a central depot and a number of retailers. In their study, the retailers satisfy customer demand with the stock kept in the central depot. The inventory is periodically reviewed and replenished. From the central depot to the retailers, stock is replenished by the share rationing policy, where the ratio of the projected net inventory at a retailer over the system-wide projected net inventory is kept constant and determined by the customer service level required by the retailer. Under the policy, when an order arrives, the total net stocks amongst the retailers are reallocated by the transshipment of stocks in order to maintain the constant ratio. The authors proposed a heuristics to obtain the ratios that minimize the total expected cost. Their numerical results show that, compared with the system without transshipment, such transshipment significantly reduces the total expected cost.

Existing studies on inventory sharing in a supply chain focus on the scenario where the inventory policies for each retailer are coordinated. Such scenario is applicable if

a single “parent firm” owns all the retailer and it tries to identify the inventory allocation that maximize the supply chain performance (usually the total profit). Rudi et al. (2001) consider transshipment among different firms, where each firm tries to maximize its own profit. Three different scenarios – without transshipment, coordinated transshipment and decentralized transshipment – are examined. While in general, the decentralized transshipment decisions do not maximize the total profit for the supply chain, there exist transshipment prices that induce the locations to maximize the total expected profit.

Apart from inventory sharing, cargoes may be consolidated in transshipment hubs, as commonly seen in postal service industries and in third party logistics providers. Cargo consolidation reduces transportation costs due to economies of scale inherent in freight transportation and distribution. Many firms have been consolidating cargoes for a number of years as an effective way to reduce transportation costs. Consolidation has also become a subject of greater interest as the difference between truckload and less-than-truckload rates has increased (Higginson and Bookbinder 1994).

The hub and spoke system (e.g., airline routes, supply chain and telecommunication networks) connect origins and destinations via one or more hubs and these hubs are connected by an efficient transportation system. For the design of the hub and spoke system, heuristic algorithms have been developed (Pirkul and Schilling 1998, Skorin-Kapov and O’Kelly 1996, Wiles and van Brunt 2001). Amongst them, Pirkul and Schilling (1998) provides an efficient heuristic algorithm with tight bounds. The computational experiments on eighty-four standard test problems show that the average gaps are 0.048%, and the maximum gap is less than 1%.

De Rosa et al. (2002) investigates an arc routing and scheduling problem with transshipment, in which goods are first collected and taken to a transshipment hub for processing and transported to the same final destination by a high-capacity vehicle. A tight lower bound is provided by the sum of a lower bound on the traversal cost of all the required edges and a lower bound on the total cost of a truck schedule. Based on the bounds, an efficient heuristics employing Tabu search techniques are then developed.

Shigeno et al. (2000) provide a comprehensive literature review on the minimum cost network flow problem, focusing on cycle and cut canceling algorithms. Marin and Pelegrin (1996) consider a transshipment problem with fixed cost for each transshipment hub. The number of hubs is assumed to be given. They decompose the problem into two sub-problems by the Lagrangean decomposition. Using a dual ascent algorithm, a lower bound for the optimal objective function value is obtained. Furthermore, an efficient heuristics algorithm is proposed. The transshipment problem considered in this paper is different since the number of transshipment hubs is not given in this study. We also consider transportation cost to be non-linear and a step-decreasing function.

3 Negotiation Process for the Supply Chain Network Design

In a supply chain network design, many companies are involved. These companies will confront problems even though they follow the rules and contracts; therefore they

need to resolve those problems together. In order to increase the effectiveness of the supply chain network design, the companies need to follow an integrated negotiation approach rather than a distributive negotiation approach. Then, they can reach a win-win solution or integrated solution.

If there is just one issue to be resolved, negotiators should approach the problem with distributive strategy. In this case it is a zero-sum game, since a negotiator's gain is the other party's loss and the other party's gain is the focal negotiator's loss. In addition, the amount of loss and gain are exactly same. In many cases, however, negotiators deal with several issues at the same time and negotiators may have different importance for the specific issues. Here, there is a potential for a win-win solution, if they can trade-off those two issues. That is, a negotiator can concede on an issue that is less important to him or her but important to the other party. At the same time the negotiator can be firm about an important issue to himself or herself but less important to the other party and ask the other party to make concession. By adopting integrative negotiation approach, they can increase the effectiveness of the total supply chain.

4 Model and Solution Approach

In this study, we consider a supply chain network that consists of nodes and arcs. A set of cities (in-between transportation of goods are required) and a set of promising transshipment hub locations comprise nodes in the network, and the nodes are connected by arcs in the network. Every pair of nodes is connected by at least one arc. When more than two arcs connect a pair of nodes, each arc represents distinctive transportation mode.

All the arcs are capacitated, i.e., there exists an upper bound for the amount of goods to transfer via each arc, and also the transshipment hubs are capacitated, i.e., there exists an upper bound for the amount of cargo each transshipment hub can handle. For each unit of cargo flowing through a transshipment hub, a fixed transshipment handling cost is charged. We further assume that the unit transportation cost of each arc follows a specific decreasing step function, i.e. when the total cargo flow on an arc exceeds certain amount, the exceeding amount will be charged a lower unit cost.

The object of our problem is to determine which of the potential transshipment hubs should be selected and what pattern of the cargo flow should follow, in order to minimize the total cost of the whole transportation system, which is the sum of the costs of transportation, of transshipment handling, and of hub lease. The mixed integer programming formulation of the problem is as follows:

The objective function is the total system cost that consists of transportation cost, transshipment handling cost and hub lease cost, respectively. The constraints set consists of flow balance constraints, capacity constraints and integrality constraints.

Due to the complexity of this problem, solving the original mixed integer programming model is very difficult. Instead, we propose a two-stage heuristic approach to solve the problem. In the first stage, we obtain the hub location, and then in the second stage, transportation decisions that include how many goods will be transported from a node to the other nodes and what transportation mode will be selected for the flow are determined.

In order to solve the first stage problem, we compute the cost increase (decrease) by removing a location from the set of transshipment hubs, one at a time. Starting with all the promising locations to be transshipment hubs, we remove the location that reduces the total cost the most at each iteration. The iterative search is terminated when there is no location whose removal may further reduce the total cost. In the second stage, based on the given transshipment hub location, we design the transportation flow by solving a multi-commodity network flow problem.

5 Implementation

This study was commissioned partly by a multi-national logistics service provider, and the solution obtained from the proposed method is then applied to a network design problem that the company is facing in a busy part of their Asia Pacific distribution network. The company analysis results show that the total cost can be reduced by more than 20% by relocating transshipment hubs and considering multi-modal transportation, as obtained by the solution procedure in this study. Furthermore, the proposed heuristic approach provides solutions redesign cargo flows in very short time, and thus, can be employed to reroute cargoes in cases of operational contingency such as 9-11.

6 Conclusions

In this study, we study a multi-modal transportation network with transshipment. Our objective is to select hub locations amongst several promising locations and design the cargo flow in the distribution network to minimize the total expected cost, which is the total of transportation costs, transshipment handling costs and fixed costs. In addition, we consider the supply chain where many companies are involved to maximize their own profits. In order for the win-win contract design that maximizes the effectiveness of the supply chain network design, the companies need to follow an integrated negotiation approach rather than a distributive negotiation approach.

The mixed integer programming model for the distribution network design problem is hard to solve. Thus, a two-stage heuristic approach to solve the mixed integer programming problem is proposed. In the first stage, we obtain the hub location, and then in the second stage, transportation decisions that include how many goods will be transported from a node to the other nodes and what transportation mode will be selected for the flow are determined. The proposed two-stage heuristic approach is applied to a real-life third party logistics provider's problem and reduces about 20% of the total system cost.

References

- Diks, E. B. and de Kok, A. G., "Controlling a Divergent 2-echelon Network with Transshipments Using the Consistent Appropriate Share Rationing Policy", *International Journal of Production Economics* 45 (1996) 369-379
- Higginson, J. and Bookbinder, J., "Policy Recommendation for a Shipment Consolidation Program", *Journal of Business Logistics* 15 (1994) 87-112

- Marin, A., and Pelegrin, B., "A Branch-and-bound Algorithm for the Transportation Problem with Location of p Transshipment Points", *Computers and Operations Research* 24 (1997) 659-678
- Pirkul, H., and Schilling, D. A., "An Efficient Procedure for Designing Single Allocation Hub and Spoke Systems", *Management Science* 44 (1998) 235-242
- Rosa, B. D., Improta, G., Ghiani, G., and Musmanno, R., "The Arc Routing and Scheduling Problem with Transshipment", *Transportation Science* 36 (2002) 301-313
- Rudi, N., Kapur, S., and Pyke, D. F., "A Two-Location Inventory Model with Transshipment and Local Decision Making", *Management Science* 47 (2001) 1668-1680
- Shigeno, M., Iwata, S., and McCormick, S. T., "Relaxed Most Negative Cycle and Most Positive Cut Canceling Algorithms for Minimum Cost Flow", *Mathematics of Operational research* 25 (2000) 76-83
- Skorin-Kapov, D. and O'Kelly M. E., "Tight linear programming relaxations of uncapacitated p -hub median problems", *European Journal of Operational Research* 94 (1996) 582-593.
- The Logistics Institute Asia Pacific, "Third - Party Logistics Results and Findings of the 2004 Ninth Annual Study", Singapore (2004)
- Viswanadham, N., Jarvis, J. J., Gaonkar, R. S., "Ten Mega Trends in Logistics", *White Paper*, The Logistics Institute Asia Pacific, 2003
- Wiles, P. G., and van Brunt, B., "Optimal Location of Transshipment Depots" *Transportation Research Part A* 35 (2001) 745-771

A Group Search Optimizer for Neural Network Training

S. He¹, Q.H. Wu^{1,*}, and J.R. Saunders²

¹ Department of Electrical Engineering and Electronics

² School of Biological Sciences,

The University of Liverpool, Liverpool, L69 3GJ, UK

qhwu@liv.ac.uk

Abstract. A novel optimization algorithm: Group Search Optimizer (GSO) [1] has been successfully developed, which is inspired by animal behavioural ecology. The algorithm is based on a Producer-Scrounger model of animal behaviour, which assumes group members search either for ‘finding’ (producer) or for ‘joining’ (scrounger) opportunities. Animal scanning mechanisms (*e.g.*, vision) are incorporated to develop the algorithm. In this paper, we apply the GSO to Artificial Neural Network (ANN) training to further investigate its applicability to real-world problems. The parameters of a 3-layer feed-forward ANN, including connection weights and bias are tuned by the GSO algorithm. Two real-world classification problems have been employed as benchmark problems trained by the ANN, to assess the performance of the GSO-trained ANN (GSOANN). In comparison with other sophisticated machine learning techniques proposed for ANN training in recent years, including some ANN ensembles, GSOANN has a better convergence and generalization performances on the two benchmark problems.

1 Introduction

Artificial Neural Networks (ANNs) have been widely applied to a variety of problem domains such as pattern recognition [2] and control [3] since their renaissance in the mid-1980’s. Various ANN architectures and training algorithms have been proposed. Among them, the most popular ANN architecture and training algorithm are feed-forward ANNs and the BP training algorithm, respectively. However, the gradient-based BP training algorithm is easy to be trapped by local minima and therefore deteriorates the performance of ANNs. On the other hand, designing a near optimal ANN architecture to achieve good generalization performance is also a hard optimization problem.

In the past two decades, Evolutionary Algorithms (EAs) have been introduced to ANNs to perform various tasks, such as connection weight training, architecture design, learning rule adaption, input feature selection, connection weight initialization, rule extraction from ANN, etc.[4]. The combinations of ANNs and EAs are usually referred to as Evolutionary ANNs (EANNs). The earliest attempt to combine EAs and ANNs can be traced back to the late 1980s.

* Corresponding author.

Since then, the successful marriage of ANNs and EAs has attracted more and more attention [5] [6]. In [7], an improved genetic algorithm was used to tune the structure and parameters of a neural network. An improved Genetic Algorithm (GA) with new genetic operators were introduced to train the proposed ANN. Two application examples, sunspots forecasting and associative memory tuning, were solved in their study.

Palmes *et al.* proposed a mutation-based genetic neural network (MGNN) [8]. A simple matrix encoding scheme was used to represent an ANN's architecture and weights. The neural network utilized a mutation strategy of local adaptation of evolutionary programming to evolve network structures and connection weights dynamically. Three classification problems, namely iris classification, wine recognition problem, and Wisconsin breast cancer diagnosis problem were used in their paper as benchmark functions.

Cantú-Paz and Kamath presented an empirical evaluation of eight combinations of EAs and ANNs on 11 well studied real-world benchmarks and 4 synthetic problems [9]. The algorithms they used included binary-encoded, real-encoded GAs, and the BP algorithm. The tasks performed by these algorithms and their combinations included searching for weights, designing architecture of ANNs, and selecting feature subsets for ANN training.

We have proposed a novel GSO for continuous optimization problems [1], it is quite logically to apply our GSO algorithm to ANN weight training. The ANN weight training process can be regarded as a hard continuous optimization problem, since the search space is high-dimensional multi-modal and is usually polluted by noises and missing data. The objective of ANN weight training process is to minimize an ANN's error function. However, it has been pointed out that minimizing the error function is different from maximizing generalization [10]. Therefore, to improve ANN's generalization performance, in this study, an early stopping scheme is introduced. The error rates of validation sets are monitored during the training processes. When the validation error increases for a specified number of iterations, the training will stop. The GSO algorithm and early stopping scheme have been applied to training an ANN for two benchmark functions - Wisconsin breast cancer data set and Pima Indian diabetes data set.

The rest of the paper is organized as follows. We present the GSO algorithm in Section 2. In Section 3, GSOANN will be introduced and the details of implementation will be given. In Section 4, we describe the benchmark functions, experimental settings and the experimental results. The paper is concluded in Section 5.

2 Group Search Optimizer

Optimization, which is a process of seeking optima in a search space, is analogous to the resource searching process of animals in nature. It is quite natural to draw inspiration from animal searching behavior, especially group searching behavior to develop an optimization algorithm. The GSO which was inspired by animal searching behavior and group living theory was proposed.

The GSO algorithm employs the Producer-Scrounger (PS) model [11] as a framework. The PS model was proposed to analyze social foraging strategies of group living animals. There are two foraging strategies within groups: (1) producing, *e.g.*, searching for food; and (2) joining (scrounging), *e.g.*, joining resources uncovered by others. In the PS model, foragers are assumed to use producing or joining strategies exclusively. Concepts of resource searching from animal scanning mechanism and the PS model are used to design optimum searching strategies for GSO. Basically GSO is a population based optimization algorithm. The population of the GSO algorithm is called a *group* and each individual in the population is called a *member*. In an n -dimensional search space, the i_{th} member at the k_{th} searching bout (iteration), has a current position $X_i^k \in \mathbb{R}^n$, a head angle $\varphi_i^k = (\varphi_{i1}^k, \dots, \varphi_{i(n-1)}^k) \in \mathbb{R}^{n-1}$ and a head direction $D_i^k(\varphi_i^k) = (d_{i1}^k, \dots, d_{in}^k) \in \mathbb{R}^n$ which can be calculated from φ_i^k via a Polar to Cartesian coordinates transformation:

$$\begin{aligned}
 d_{i1}^k &= \prod_{p=1}^{n-1} \cos(\varphi_{ip}^k) \\
 d_{ij}^k &= \sin(\varphi_{i(j-1)}^k) \cdot \prod_{p=i}^{n-1} \cos(\varphi_{ip}^k) \\
 d_{in}^k &= \sin(\varphi_{i(n-1)}^k)
 \end{aligned} \tag{1}$$

In the GSO, a group comprises three kinds of members: producers, scroungers and rangers. The behaviors of producers and scroungers are based on the PS model. We also employ ‘rangers’ which perform random walks to avoid entrapment in local minima. For accuracy [12] and convenience of computation, in the GSO algorithm, there is only one producer at each searching bout and the remaining members are scroungers and rangers. The simplest joining policy, which assumes all scroungers will join the resource found by the producer, is used.

During each search bout, a group member, located in the most promising area, conferring the best fitness value, acts as the producer. It then stops and scans the environment to search resources (optima). Scanning can be accomplished through physical contact or by visual, chemical, or auditory mechanisms [13]. Vision, as the main scanning mechanism used by many animal species, is employed by the producer in GSO. In order to handle optimization problems of whose number of dimensions usually is larger than 3, the scanning field of vision is generalized to a n dimensional space, which is characterized by maximum pursuit angle $\theta_{\max} \in \mathbb{R}^{n-1}$ and maximum pursuit distance $l_{\max} \in \mathbb{R}^1$ as illustrated in a 3D space in Figure 1. In the GSO algorithm, at the k_{th} iteration the producer X_p behaves as follows:

- 1) The producer will scan at zero degree and then scan laterally by randomly sampling three points in the scanning field [14]: one point at zero degree:

$$X_z = X_p^k + r_1 l_{\max} D_p^k(\varphi^k) \tag{2}$$

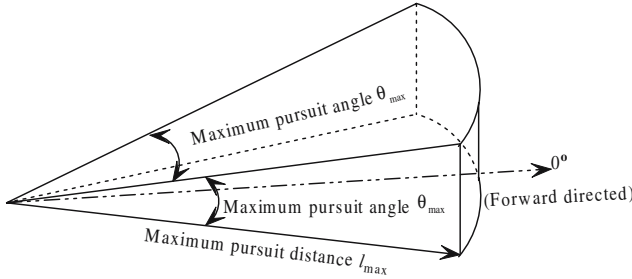


Fig. 1. Scanning field in 3D space [13]

one point in the right hand side hypercube:

$$X_r = X_p^k + r_1 l_{\max} D_p^k (\varphi^k + r_2 \theta_{\max} / 2) \tag{3}$$

and one point in the left hand side hypercube:

$$X_l = X_p^k + r_1 l_{\max} D_p^k (\varphi^k - r_2 \theta_{\max} / 2) \tag{4}$$

where $r_1 \in \mathbb{R}^1$ is a normally distributed random number with mean 0 and standard deviation 1 and $r_2 \in \mathbb{R}^{n-1}$ is a random sequence in the range (0, 1).

- 2) The producer will then find the best point with the best resource (fitness value). If the best point has a better resource than its current position, then it will fly to this point. Otherwise it will stay in its current position and turn its head to a new angle:

$$\varphi^{k+1} = \varphi^k + r_2 \alpha_{\max} \tag{5}$$

where α_{\max} is the maximum turning angle.

- 3) If the producer cannot find a better area after a iterations, it will turn its head back to zero degree:

$$\varphi^{k+a} = \varphi^k \tag{6}$$

where a is a constant given by $\text{round}(\sqrt{n+1})$.

At each iteration, a number of group members are selected as scroungers. The scroungers will keep searching for opportunities to join the resources found by the producer. The commonest scrounging behavior [11] in house sparrows (*Passer domesticus*): area copying, that is, moving across to search in the immediate area around the producer, is adopted. At the k th iteration, the area copying behavior of the i th scrounger can be modeled as a random walk towards the producer:

$$X_i^{k+1} = X_i^k + r_3 (X_p^k - X_i^k) \tag{7}$$

where $r_3 \in \mathbb{R}^n$ is a uniform random sequence in the range (0, 1).

In nature, group members often have different searching and competitive abilities; subordinates, who are less efficient foragers than the dominant will be

dispersed from the group [15]. This may result in ranging behavior. Ranging is an initial phase of a search that starts without cues leading to a specific resource [16]. The ranging animals - rangers, may explore and colonize new habitats. In our GSO algorithm, rangers are introduced to explore a new search space therefore to avoid entrapments of local minima. The rangers perform search strategies which include random walks and systematic search strategies to locate resources efficiently [17]. In the GSO algorithm, random walks, which are thought to be the most efficient searching method for randomly distributed resources [18], are employed by rangers. If the i_{th} group member is selected as a ranger, at the k_{th} iteration, it generates a random head angle φ_i :

$$\varphi_i^{k+1} = \varphi_i^k + r_2 \alpha_{\max} \quad (8)$$

where α_{\max} is the maximum turning angle; and it chooses a random distance:

$$l_i = a \cdot r_1 l_{\max} \quad (9)$$

and move to the new point:

$$X_i^{k+1} = X_i^k + l_i D_i^k(\varphi^{k+1}) \quad (10)$$

In order to maximize their chances of finding resources, animals restrict their search to a profitable patch. One strategy is turning back into a patch when its edge is detected [19]. This strategy is employed by GSO to handle the bounded search space: when a member is outside the search space, it will turn back to its previous position inside the search space.

3 The GSO Based Training Algorithm for Neural Networks

Figure 2 presents a three-layer feed-forward ANN to be tuned by our GSO algorithm. The ANN consists three layers, namely, input, hidden, and output layers. The nodes in each layer receive input signals from the previous layer and pass the output to the subsequent layer. The nodes of the input layer supply respective elements of the activation pattern (input vector), which constitute the input signals from outside system applied to the nodes in the hidden layer by the weighted links. The output signals of the nodes in the output layer of the network constitute the overall response of the network to the activation pattern supplied by the source nodes in the input layer [20]. The subscripts n , h , and k denote any node in the input, hidden, and output layers, respectively. The net input u is defined as the weighted sum of the incoming signal minus a bias term. The net input of node h , u_h , in the hidden layer is expressed as follows:

$$u_h = \sum^n w_{hn} y_n - \theta_h$$

where y_n is the output of node n in the input layer, w_{hn} represents the connection weight from node n in the input layer to node h in the hidden layer, and θ_h is the

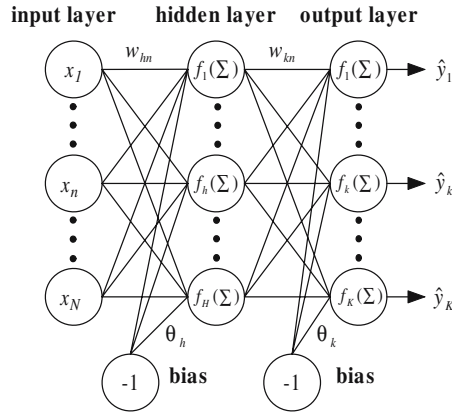


Fig. 2. A three-layer feed-forward ANN

bias of node h in the hidden layer. The activation function used in the proposed ANN is the sigmoid function. Therefore, in the hidden layer, the output y_h of node h , can be expressed as

$$y_h = f_h(u_h) = \frac{1}{1 + e^{u_h}}$$

The output of node k in the output layer can be also described as:

$$y_k = f_k(u_k) = \frac{1}{1 + e^{u_k}} \tag{11}$$

where

$$u_k = \sum_h w_{kh} y_h - \theta_k$$

where θ_k is the bias of node k in the output layer.

The parameters (connection weights and bias terms) are tuned by the GSO algorithm. In the GSO-based training algorithm, each member of the population is a vector comprising connection weights and bias terms. Without loss of generality, we denote W_1 as the connection weight matrix between the input layer and the hidden layer, Θ_1 as the bias terms to the hidden layer, W_2 as the one between the hidden layer and the output layer, and Θ_2 as the bias terms to the output layer. The i_{th} member in the population can be represented as: $X_i = [W_1^i \ \Theta_1^i \ W_2^i \ \Theta_2^i]$. The fitness function assigned to the i_{th} individual is the least-squared error function defined as follows:

$$F_i = \frac{1}{2} \sum_{p=1}^P \sum_{k=1}^K (d_{kp} - y_{kp}^i)^2 \tag{12}$$

where y_{kp}^i indicates the k_{th} computed output in equation (11) of the ANN for the p_{th} sample vector of the i_{th} member; P denotes the total number of sample vectors; and d_{kp} is the desired output in the k_{th} output node.

4 Experimental Studies

In order to evaluate the GSOANN's performance, two well-studied benchmark problems from the UCI machine learning repository were tested. These problems are Wisconsin breast classification data and Pima Indian diabetes data. They are all real-world problems which are solved by human experts in practice. The data sets of these problems usually contain missing attribute values and are polluted by noise. Therefore, they represent some of the most challenging problems in the machine learning field [6]. We evaluate the GSO algorithm on these two problems and compare our results with the latest results published in the literature.

4.1 Experimental Setting

The parameter setting of the GSO algorithm is as follows. The initial population of GSO is generated at uniformly random in the search space. The initial head angle φ^0 of each individual is set to be $\frac{\pi}{4}$. The maximum pursuit angle θ_{\max} is $\frac{\pi}{a^2}$. The maximum turning angle α is set to be $\frac{\pi}{2a^2}$. The maximum pursuit distance l_{\max} is calculated from:

$$l_{\max} = \| U_i - L_i \| = \sqrt{\sum_{i=1}^n (U_i - L_i)^2}$$

where L_i and U_i are the lower and upper bounds for the i_{th} dimension. The parameter which needs to tune is the percentage of rangers; our recommended percentage of rangers is 20%, which was used throughout all our experiments. The population size of the GSO algorithm was set to 50.

The data sets of the four classification problems were partitioned according the guidelines of Prechelt [21]. Each set of data was divided into three sets: 50% of the patterns were used for learning, 25% of them for validation and the remaining 25% for testing the generalization of the trained ANN. The maximum epoches varied according to the convergence rates of the training algorithms on different problems. Several runs were executed to determine the maximum epoches for the five benchmark problems. All experiments were repeated 30 runs in order to get average results. The proposed algorithm and the other six training algorithms were implemented in MATLAB 6.5 and executed on a Pentium 4, 2.0 GHz machine.

4.2 The Wisconsin Breast Cancer Data Set

The breast cancer data set was obtained by W. H. Wolberg *et al.* at the University of Wisconsin Hospitals, Madison. The data set currently contains 9 integer-valued attributes and 699 instances of which 458 are benign and 241 are malignant example. In order to train ANNs to classify a tumor as either benign

or malignant, we partitioned this data set into three sets: a training set which contains the first 349 examples, a validation set which contains the following 175 examples, and a test set which contains the final 175 examples.

The comparisons between the results produced by GSOANN and those of 9 other machine learning algorithms are tabulated in Table 1. Among these algorithms, MGNN [8] and EPNet [6] evolved ANN structure as well as connection weights; COOP [22] is an evolutionary ANN ensemble evolved by cooperative coevolution; CNNE [23] is a constructive algorithm for training cooperative ANN ensembles. CCSS [24], OC1-best [25] and EDTs [26] are state-of-the-art decision tree classifiers, including the decision tree ensembles [24] [26] and hybrid evolutionary decision tree [25]; GANet-best is the best result from [9], which was generated by an EANN based on a real-encoded EA [27] to evolve connection weights; the SVM-best is the best result of 8 least squares SVM classifiers [28]. It is worth to mention that the decision trees [24] [26] [9] and SVM [28] techniques used k -fold cross-validation which generated more optimistic results.

In comparison with the sophisticated classifiers mentioned above, we can find that this simple GSOANN produced the best average result.

Table 1. Comparison between GSOANN and other approaches in terms of average testing error rate (%) on the Wisconsin breast cancer data set

Algorithm	GSOANN	GANet-best[9]	COOP [22]	CNNE [23]	EPNet [6]
Test error rate (%)	0.65	1.06	1.23	1.20	1.38
Algorithm	MGNN [8]	SVM-best [28]	CCSS [24]	OC1-best [25]	EDTs [26]
Test error rate (%)	3.05	3.1	2.72	3.9	2.63

4.3 The Pima Indian Diabetes Data Set

The Pima Indian diabetes data was originally donated by Vincent Sigillito at the Johns Hopkins University. The diagnostic, binary-valued variable investigated is whether a patient shows signs of diabetes according to World Health Organization criteria. There are 8 numeric-valued attributes and 768 instances. The data set contains 500 instances of patients with signs of diabetes and 268 instances of patients without. The data set were partitioned: the first 384 instances were used as the training set, the following 192 instances as the validation set, and the final 192 instances as the test set.

This problem is one of the most difficult machine learning problems since the data set is relatively small and was heavily polluted by noise. Results from other state-of-the-art classifiers are tabulated in Table 2. COVNET [29] is a cooperative coevolutionary model for evolving artificial neural networks. EENCL is evolutionary ensembles with negative correlation learning presented in [30]. Twelve-fold cross-validation was used by EENCL. The GANet-best is the best result produced by an ANN trained by a subset of features selected by a binary-encoded GA [9].

From Table 2, it can be seen that GSOANN is outperformed by COOP [22] and CNNE [23] which are both ANN ensembles. However, GSOANN produced

Table 2. Comparison between GSOANN and other approaches in terms of average testing error rate (%) on the Pima diabetes disease data set

Algorithm	GSOANN	GANet-best[9]	COOP[22]	CNNE[23]	COVNET[29]
Test error rate (%)	19.79	24.70	19.69	19.60	19.90
Algorithm	EENCL [30]	EPNet [6]	SVM-best [28]	CCSS [24]	OC1-best [25]
Test error rate (%)	22.1	22.38	22.7	24.02	26.0

better results than those of the rest classifiers including evolutionary ANN ensembles COVNET [29] and EENCL [30].

5 Conclusion

In this paper, our GSO algorithm has been applied to train ANN's connection weights. Our initial goal was not to propose a sophisticated ANN which can achieve the best generalization performance. Instead, we aimed to access GSO's global search performance on real-world problems by applying it for ANN training since the training process can be regarded as a hard continuous optimization problem. Our experimental results show that, it is not enough for solving the real-world classification problems to compare the results of GSOANN with those of other sophisticated ANNs such as ANN ensembles. However, since the GSOANN is relatively simple, only on the aspect of ANN's weight training, this GSO based ANN training algorithm provides relatively better results on the two benchmark problems we tested.

References

1. He, S., Wu, Q.H., Saunders, J.R.: Group search optimizer - an optimization algorithm inspired by animal behavioral ecology. (Submitted to IEEE Trans. on Evolutionary Computation)
2. Thrun, S.B., et. al.: The MONK's problems: A performance comparison of different learning algorithms. Technical Report CS-91-197, Pittsburgh, PA (1991)
3. Wu, Q.H., Hogg, B.W., Irwin, G.W.: A neural network regulator for turbogenerators. IEEE Trans. on Neural Networks **3**(1) (1992) 95–100
4. Yao, X.: Evolving artificial neural networks. Proceeding of the IEEE **87**(9) (1999) 1423–1447
5. Fogel, D.B., Fogel, L.J., Porto, V.W.: Evolving neural networks. Biol. Cybern. **63** (1990) 487–493
6. Yao, X., Liu, Y.: A new evolutionary system for evolving artificial neural networks. IEEE Trans. on Neural Networks **8**(3) (1997) 694–713
7. Leung, F.H.F., Lam, H.K., Ling, S.H., Tam, P.K.S.: Tuning of the structure and parameters of a neural network using an improved genetic algorithm. IEEE Trans. on Neural Networks **14**(1) (2003) 79–88
8. Palmes, P.P., Hayasaka, T., Usui, S.: Mutation-based genetic neural network. IEEE Trans. on Neural Networks **16**(3) (2005) 587–600

9. Cantu-Paz, E., Kamath, C.: An empirical comparison of combinations of evolutionary algorithms and neural networks for classification problems. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics* **35**(5) (2005) 915–927
10. Wolpert, D.H.: A mathematical theory of generalization. *Complex Systems* **4**(2) (1990) 151–249
11. Barnard, C.J., Sibly, R.M.: Producers and scroungers: a general model and its application to captive flocks of house sparrows. *Animal Behaviour* **29** (1981) 543–550
12. Couzin, I., Krause, J., Franks, N., Levin, S.: Effective leadership and decision-making in animal groups on the move. *Nature* **434** (2005) 513–516
13. Bell, J.W.: *Searching Behaviour - The Behavioural Ecology of Finding Resources*. Chapman and Hall Animal Behaviour Series. Chapman and Hall (1990)
14. O'Brien, W.J., Evans, B.I., Howick, G.L.: A new view of the predation cycle of a planktivorous fish, white crappie (*pomoxis annularis*). *Can. J. Fish. Aquat. Sci.* **43** (1986) 1894–1899
15. Harper, D.G.C.: Competitive foraging in mallards: 'ideal free' ducks. *Animal Behaviour* **30** (1988) 575–584
16. Dusenbery, D.B.: Ranging strategies. *Journal of Theoretical Biology* **136** (1989) 309–316
17. Higgins, C.L., Strauss, R.E.: Discrimination and classification of foraging paths produced by search-tactic models. *Behavioral Ecology* **15**(2) (2003) 248–254
18. Viswanathan, G.M., Buldyrev, S.V., Havlin, S., da Luz, M.G., Raposo, E., Stanley, H.E.: Optimizing the success of random searches. *Nature* **401**(911-914) (1999)
19. Dixon, A.F.G.: An experimental study of the searching behaviour of the predatory coccinellid beetle *adalia decempunctata*. *J. Anim. Ecol.* **28** (1959) 259–281
20. Haykin, S.: *Neural Networks*. A Comprehensive Foundation. Prentice Hall, New Jersey, USA (1999)
21. Prechelt, L.: *Problem1 - a set of neural network benchmark problems and benchmarking rules*. Technical report, Fakultat fur Informatik Universitat Karlsruhe, 76128 Karlsruhe, Germany (1995)
22. Garcia-Pedrajas, N., Hervas-Martinez, C., Ortiz-Boyer, D.: Cooperative coevolution of artificial neural network ensembles for pattern classification. *IEEE Trans. on Evolutionary Computation* **9**(3) (2005) 271–302
23. Islam, M., Yao, X., Murase, K.: A constructive algorithm for training cooperative neural network ensembles. *IEEE Trans. on Neural Networks* **14**(4) (2003) 820–834
24. Dzeroski, S., Zenko, B.: Is combining classifiers with stacking better than selecting the best one? *Machine Learning* **54**(3) (2004) 255–273
25. Cantu-Paz, E., Kamath, C.: Inducing oblique decision trees with evolutionary algorithms. *IEEE Trans. on Evolutionary Computation* **7**(1) (2003) 54–68
26. Ditzterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning* **40**(12) (2000) 139–157
27. Deb, K., Anand, A., Joshi, D.: A computationally efficient evolutionary algorithm for real-parameter optimization. *Evolutionary Computation* **10**(4) (2002) 371–395
28. Gestel, T.V., et. al.: Benchmarking least squares support vector machine classifiers. *Machine Learning* **54**(1) (2004) 5–32
29. Garcia-Pedrajas, N., Hervas-Martinez, C., Munoz-Perez, J.: Covnet: a cooperative coevolutionary model for evolving artificial neural networks. *IEEE Trans. on Neural Networks* **14**(3) (2003) 575–596
30. Liu, Y., Yao, X.: Evolutionary ensembles with negative correlation learning. *IEEE Trans. on Evolutionary Computation* **4**(4) (2000) 380–387

Application of Two-Stage Stochastic Linear Program for Portfolio Selection Problem

Kuo-Hwa Chang, Huifen Chen, and Ching-Fen Lin

Department of Industrial Engineering, Chung Yuan Christian University,
Chung-Li, 320, Taiwan

Abstract. We consider a portfolio selection problem under the consideration of dynamic closing time of the portfolio. The selection strategy is to take the long position on the stocks and the short position on an index future. We will close our portfolio whenever our profit exceeds the predetermined target during the investment period, otherwise, we will own the portfolio till the maturity date of the future. Our purpose is to have a profitable portfolio with steady return which is higher than the interest rate of savings and independent of the market. To deal with the stocks selection problem and, at the same time, the uncertainty on the closing time due to the fluctuation of the market, we define the corresponding optimization problem as a two-stage stochastic linear program (two-stage SLP). Our models are tested by the real-world data and the results are consistent with what we expected.

Keywords: Portfolio selection, Futures, Two-stage stochastic linear program.

1 Introduction

Portfolio theory deals with the problem how to allocate wealth among several assets. The basic theory of portfolio optimization was presented by Markowitz [11]. By employing the standard deviation and expected value of the asset as the parameters, Markowitz introduced the famous mean-variance (MV) model. There has been a tremendous amount of researches on improving this basic model. Kane [6] considers skewness in the model additionally, called MVS(mean-variance-skewness) model. Single index model was another improved model. By assuming a linear relation between the return of assets and the return of the market index, the single-index model reduces the amount of input data. Safety-first models consider limiting the risk of bad outcomes. Telser [12] proposes safety-first a model to maximize expected return, subjected to the constraint that the probability of a return less than, or equal to, some predetermined limit will not greater than some predetermined number.

Because of the computational difficulty in solving a large-scale MV model, Konno and Yamazaki [8] defines mean-absolute deviation (MAD) to replace covariance as the measure of risk. The major advantage of the MAD model is that it can be solved as a linear program instead of a quadratic program for

an MV model. Konno and Kobayashi [7] constructs an integrated stock-bond portfolio optimization model by minimizing the absolute deviation of the return rate of the portfolio with a given expected return rate. Cai et al. [1] provides a portfolio selection rule whose objective was to minimize the maximum individual risk and the MAD was used as the risk measure. Konno et al. [9] further solves an MVS model as a linear program.

Some risk-aversion investor may want to have a portfolio whose return rate is relatively stable comparing with market but is higher than saving interest rate. Financial derivatives such as futures and options help investors to divert and further to hedge their investment risks. An intuitive portfolio would be a bucket of stocks taken in long position and an index future in short position. Chang [3] studies this kind of portfolio and the holding time of this portfolio starts from the first date when the future is issued and ends on the maturity date of this future. In this paper, we further add more flexible investment strategy that we can close all positions during the investment period once the profit exceeds the predetermined target. However, under the uncertainty of the closing time, we should be able to find an alternative optimization model to deal with the uncertainty. In this paper, we model the proposed portfolio problem as a two-stage stochastic linear programming (SLP) problem. Stochastic program is one of the most widely applicable models incorporating uncertainty within optimization objective. It is used in many fields, such as production planning. It is also used for solving some financial planning models, especially the asset liability management (ALM). For example, Kouwenberg [10] presents a scenario generation method for solving the stochastic program associated with an ALM model.

Two-stage SLP provides a suitable framework for modeling decision problems that incorporate uncertainty and, most importantly, it can be solved by using simplex method. Chang et al. [2] extends MAD model to a two-stage SLP model for the portfolio selection problem with transaction costs. In their model, investment decision is made in the first stage and the uncertainty from the return rates is handled by the second stage. Our purpose is, under the consideration that we can close our portfolio once our profit exceeds our profit target or we will close it on the maturity date of the future, to determine an optimal portfolio with stocks in long positions and an index future in short position. The uncertainty comes from the closing time. We model this problem as a two-stage SLP and solve it by the corresponding decomposition procedure.

Let x be the column vector representing decision variables for the main problem (stage 1) and y be the one for the stage 2. The general framework of two-stage SLP can be formulated as follows:

$$\begin{aligned} \text{Min } & cx + E [h (x, \omega)] && \text{(Stage 1)} \\ \text{s.t. } & Ax = b \\ & x \geq 0 \end{aligned}$$

where

$$h(x, \omega) = \text{Min } dy \quad \text{(Stage 2)}$$

$$\begin{aligned} \text{s.t. } Wy &= r(\omega) - T(\omega)x \\ y &\geq 0 \end{aligned}$$

and ω' s are defined on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$. The function $E[h(x, \omega)]$ is often referred as the recourse function. In the above formulation, $Ax = b$ represents the linear constraints in stage 1 and $Wy = r(\omega) - T(\omega)x$ represents those in stage 2, in which $T(\omega)x$ accounts for the uncertainty associated with the feasible solution x .

Given x and ω , let the dual problem of stage 2 denoted by $h'(x, \omega) = \text{Max}\{g|\pi W \leq d\}$, where $g = \pi(r(\omega) - T(\omega)x)$ is the objective function. Let $X = \{x|Ax = b, x \geq 0\}$ denote the set of feasible solutions of the first stage and let the set $\Pi = \{\pi|\pi W \leq d\}$ denote the set of dual feasible solutions of the second stage problem (S). The two-stage SLP are solved by stochastic decomposition procedure(Higle and Sen [5]). The purpose of the decomposition procedure is to approximate $E[h(x, \omega)]$ by a sequence of cutting planes, which are generated by optimal dual solutions whose average values can be used as bounds for $E[h(x, \omega)]$. Let V_k denote the set of the solutions for the dual of the second stage for the first k realizations. The basic stochastic decomposition algorithm may be stated as follows. See Higle and Sen [5] for more details.

- (0) $k \leftarrow 0$. Let V_0 be an empty set and $\eta_0(x) = -\infty$. Choose an initial solution x^1 from X for the first stage, and a lower bound L .
- (1) $k \leftarrow k + 1$. Randomly generate an observation of ω^k , the k th realization, independent of any previously generated observations.
- (2) Determine $\eta_k(x)$, a piecewise linear approximation of $E[h(x, \omega)]$ by the following steps:
 - a) Solve the dual problem of the second stage for the k th realization with the objective function $g(\pi, \omega^k, x^k)$. Then obtain the optimal solution $\pi^k = \text{argmax}\{g(\pi, \omega^k, x^k)|\pi \in \Pi\}$. Update $V_k = V_{k-1} \cup \pi^k$.
 - b) Determine the coefficients of the k th cutting plane: $\alpha_k^k + \beta_k^k x = \frac{1}{k} \sum_{i=1}^k g(\pi_i^k, \omega^i, x)$, where $\pi_i^k = \text{argmax}\{g(\pi, \omega^i, x^k)|\pi \in V_k\}$ for $i = 1, \dots, k-1$.
 - c) Update the coefficients of all previous generated cuts for $i = 1, \dots, k-1$: $\alpha_i^k = \frac{k-1}{k} \alpha_i^{k-1} + \frac{1}{k} L$ and $\beta_i^k = \frac{k-1}{k} \beta_i^{k-1}$.
 - d) Let $\eta_k(x) = \text{Max}\{\alpha_i^k + \beta_i^k x | i = 1, \dots, k\}$.
- (3) Let x^{k+1} be the optimal solution of the following linear program.

$$\begin{aligned} \text{Min } cx + \eta_k \\ \text{s.t. } Ax &= b \\ \alpha_i^k + \beta_i^k x &\leq \eta_k(x) \\ x &\geq 0 \end{aligned}$$

Repeat step (1) until the sequence x^1, x^2, \dots converges.

- (4) The limiting x^k is the optimal solution.

The rest of this paper is organized as follows. In Section 2, we model our problem as a two-stage SLP. There are two models: one is with a fixed target and the other one is with a dynamic target. Both dual problems of the second stages are presented. In section 3, we test our model by considering the stocks in Taiwan Stock Exchange and the index future in Taiwan Future Exchange. We conclude our study in Section 4.

2 Portfolio Optimization Model

Assume there are n assets we can choose from. Let R_i denote the random variable representing the return rate of asset i with mean r_i and let σ_{ij} be its covariance with asset j . Let R_L denote the lowest return rate we can tolerate. In the primary model, let x_i be decision variable representing the fraction of the fund invested on asset i . By adopting Telser's safety-first model and assuming the normalities of the returns as usual, we have the following primary optimization problem

$$\begin{aligned}
 &Max \sum_{i=1}^n r_i x_i \\
 &s.t. \sum_{i=1}^n r_i x_i \geq R_L + z_\alpha \sqrt{\sum_{i=1}^n \sum_{j=1}^n \sigma_{ij} x_i x_j} \\
 &\sum_{i=1}^n x_i = 1, \quad x_i \geq 0 \quad i = 1, \dots, n,
 \end{aligned}$$

where z_α is the critical point for the standard normal distribution such that $P(X \geq z_\alpha) = \alpha$. The first constraint presents the requirement that the probability of a return greater than or equal to R_L should not be less than $1 - \alpha$.

Due to the computational difficulty associated with solving such problem with a dense covariance matrix, we use the mean-absolute deviation (MAD) to replace covariances of the return rates. MAD transforms the portfolio selection problem from a quadratic program into a linear program. Assume that we have M samples of R_i . Let r_{im} be the m th sample of R_i and let \bar{r}_i represent the corresponding average return, $\bar{r}_i = (1/M)(\sum_{m=1}^M r_{im})$. Then the mean absolute deviation is defined as $(1/M)(\sum_{m=1}^M |\sum_{i=1}^n (r_{im} - \bar{r}_i)x_i|)$ and it is used to replace the term $\sqrt{\sum_{i=1}^n \sum_{j=1}^n \sigma_{ij} x_i x_j}$. Our investment strategy is to take the long position on stocks and the short position on the index future which is used to reduce the investment risk. There are n stocks. g is the price of one index of the index future. Assume that we issue the portfolio at time zero ($t=0$). Let x_i be the number of units of the stock i will be purchased at price S_{i0} at time $t = 0$. Let F_0 be the index level of the future at time $t = 0$. Let r_{im} and r_{Fm} be the respective return rate of stock i and future sampled from past period m , $m = 1, \dots, M$ and let \bar{r}_i and \bar{r}_F be the corresponding average returns. The portfolio selection problem above with MAD approximation can be equivalently formulated as

$$\begin{aligned}
 &Max \sum_{i=1}^n \bar{r}_i x_i S_{i0} - g \bar{r}_F F_0 \\
 &s.t. \sum_{i=1}^n \bar{r}_i x_i S_{i0} - g \bar{r}_F F_0 \geq C + z_\alpha \frac{1}{M} \sum_{m=1}^M \left| \sum_{i=1}^n (r_{im} - \bar{r}_i) x_i S_{i0} - g(r_{Fm} - \bar{r}_F) F_0 \right| \\
 &\sum_{i=1}^n x_i S_{i0} \leq B, \quad x_i \geq 0 \quad i = 1, \dots, n,
 \end{aligned}$$

where C is the minimum profit we would like to obtain and B is the total budget for our investment.

As mentioned in the previous section, we further consider that we will close our portfolio once our profit exceeds the predetermined target l during the investment period, otherwise, we will have the portfolio till the maturity date of the future. To deal with the uncertainty of the closing time due to the uncertain price of each stock in the coming period, we model it as a two-stage stochastic linear problem as follows.

$$\begin{aligned}
 &Min - \sum_{i=1}^n \bar{r}_i x_i S_{i0} + g \bar{r}_F F_0 + E [h(x, \omega)] \\
 &s.t. \sum_{i=1}^n \bar{r}_i x_i S_{i0} - g \bar{r}_F F_0 \geq C + z_\alpha \frac{1}{M} \sum_{m=1}^M \left| \sum_{i=1}^n (r_{im} - \bar{r}_i) x_i S_{i0} - g(r_{Fm} - \bar{r}_F) F_0 \right| \\
 &\sum_{i=1}^n x_i S_{i0} \leq B, \quad x_i \geq 0 \quad i = 1, \dots, n,
 \end{aligned}$$

where

$$\begin{aligned}
 h(x, \omega) &= Min \left(- \sum_{t=1}^{T-1} q_t y_t \right) + (1 - \bar{y}_T) \left(\sum_{i=1}^n \bar{r}_i x_i S_{i0} - g \bar{r}_F F_0 - l \right) \\
 &s.t. \quad q_k \sum_{t=1}^k y_t \geq q_k \quad k = 1, \dots, T - 1 \\
 &\quad q_t y_t \geq 0 \quad t = 1, \dots, T - 1 \\
 &\quad \sum_{t=1}^{T-1} y_t + \bar{y}_T = 1 \\
 &\quad y_t \geq 0, \quad t = 1, \dots, T - 1, \quad \bar{y}_T \geq 0, \quad y_t \text{ and } \bar{y}_T \text{ are integers,}
 \end{aligned} \tag{P1}$$

where q_t represents the surplus profit over the target at time t and T is the length of the investment period.

In the first stage of this SLP, a portfolio is obtained under the safety-first criterion; in the second stage, the surplus profit of this given portfolio if it is closed during the investment period is obtained. For solving the problem in the second stage, we simulate those prices and the index future by using single index

model(see Elton and Gruber [4]). Let simulated $S_{it}(\omega)$ and $F_t(\omega)$ denote the price of the stock i and the price of the index future at time t during the investment period, then, for $t = 1, \dots, T - 1$, $q_t = \sum_{i=1}^n x_i(S_{it}(\omega) - S_{i0}) - g(F_t(\omega) - F_0) - l$.

In the second stage, y_t is a binary variable that $y_t = 1$ if and only if we close our portfolio at period t (once q_t is greater than zero). If there is no q_t greater than zero, the variable \bar{y}_T would be one. This means that all positions are closed till the maturity date. These constraints in the second stage will let $y_t = 1$ for the first $q_t > 0$ if there is any, where t starts from 1 to $T - 1$ and let $\bar{y}_T = 1$ otherwise. Note that the first constraint is the key one that it will assign an 1 to the y_t corresponding to the first $q_t > 0$. Also note that $\bar{y}_T = 1$ means that we will close our portfolio at the last day(maturity date) and it will imply $E [h(x, \omega)] = 0$. We define $d_m^+ - d_m^- = \sum_{i=1}^n (r_{im} - \bar{r}_i)\alpha_i - g(r_{Fm} - \bar{r}_F)F_0$ for each m to replace the absolute-value terms in the first constraint. The corresponding dual problem of (P_1) is

$$\begin{aligned}
 &Max \left(\sum_{t=1}^{T-1} q_t \pi_t \right) + \pi_T^- + \left(\sum_{i=1}^n \bar{r}_i x_i S_{i0} - g \bar{r}_F F_0 - l \right) \\
 &s.t. \sum_{t=k}^{T-1} q_t \pi_t + q_k \pi'_k + \bar{\pi}_T \leq -q_k \quad k = 1, \dots, T - 1 \\
 &\bar{\pi}_T \leq - \left(\sum_{i=1}^n \bar{r}_i x_i S_{i0} - g \bar{r}_F F_0 - l \right) \tag{D1} \\
 &\pi_t \geq 0, \quad \pi'_t \geq 0 \quad t = 1, \dots, T - 1, \bar{\pi}_T \in \mathcal{R},
 \end{aligned}$$

where $\pi = (\pi_1, \pi_2, \dots, \pi_{T-1}, \pi'_1, \pi'_2, \dots, \pi'_{T-1}, \bar{\pi}_T)$ is the row vector representing the dual variables.

Note that the target l in the above model is fixed all the time. We can further choose l as a dynamic one depending on expected performance of the coming period. Here we choose the expected revenue of the portfolio estimated in the first stage as our dynamic target. That is, $l = \sum_{i=1}^n \bar{r}_i x_i S_{i0} - g \bar{r}_F F_0$. The corresponding second stage with dynamic l is

$$\begin{aligned}
 h(x, \omega) = &Min \left(- \sum_{t=1}^{T-1} q_t y_t \right) \\
 &s.t. \quad q_k \sum_{t=1}^k y_t \geq q_k \quad k = 1, \dots, T - 1 \\
 &\sum_{t=1}^{T-1} y_t + \bar{y}_T = 1 \tag{P2} \\
 &y_t \geq 0, \quad t = 1, \dots, T - 1, \quad \bar{y}_T \geq 0, \quad y_t \text{ and } \bar{y}_T \text{ are integers.}
 \end{aligned}$$

In this model, $q_t = \sum_{i=1}^n x_i(S_{it}(\omega) - S_{i0}) - g(F_t(\omega) - F_0) - (\sum_{i=1}^n \bar{r}_i x_i S_{i0} - g \bar{r}_F F_0)$. The corresponding dual problem of (P_2) is

$$\begin{aligned}
 &Max \sum_{t=1}^{T-1} q_t \pi_t + \bar{\pi}_T \\
 &s.t. \sum_{t=k}^{T-1} q_t \pi_t + \bar{\pi}_T \leq -q_k \quad k = 1, \dots, T-1 \\
 &\bar{\pi}_T \leq 0 \\
 &\pi_t \geq 0 \quad t = 1, \dots, T-1, \quad \bar{\pi}_T \in \mathcal{R},
 \end{aligned} \tag{D_2}$$

where $\pi = (\pi_1, \pi_2, \dots, \pi_{T-1}, \bar{\pi}_T)$ is the row vector representing the dual variables.

3 Empirical Results

In this section, we test our models by considering the stocks selected from Taiwan Stock Exchange(TAIEX) and two index futures, TAIEX Futures (TX) and Mini TAIEX Futures (MTX), from Taiwan Futures Exchange. The price for one unit index for TX is NT\$200(200 New Taiwan Dollars) and NT\$ 50 for MTX. Historical data starting from 1988 up to the current month are used to estimate the parameters in our model for the coming month, based on which, the

Table 1. Return rates of Model-1 and Model-2 with TX

Month	Model-1/TX	Model-2/TX (target l)	market
Nov-03	0.0295	-0.0046 (85159)	-0.0099
Dec-03	0.0210	0.0212 (83468)	-0.0194
Jan-04	0.0097	0.0562 (75808)	0.1100
Feb-04	0.0341	-0.0114 (82780)	0.0346
Mar-04	0.0118	0.0259 (54022)	-0.0042
Apr-04	0.0328	0.0391 (42551)	0.0353
May-04	0.0441	0.0533 (44077)	-0.1394
Jun-04	0.0373	0.0482 (54973)	-0.0513
Jul-04	0.0415	0.0551 (52566)	-0.0272
Aug-04	0.0156	0.0476 (51215)	0.0034
Sep-04	0.0313	0.0401 (68083)	0.0817
Oct-04	0.0281	0.0288 (62914)	-0.0141
Nov-04	-0.0296	-0.0250 (41985)	0.0415
Dec-04	-0.0161	-0.0038 (40622)	-0.0043
Jan-05	0.0115	0.0187 (48607)	-0.0179
Feb-05	-0.0481	-0.0408 (48855)	0.0421
Mar-05	-0.0104	-0.0091 (34933)	-0.0116
Apr-05	-0.0432	-0.0412 (32330)	-0.0625
May-05	0.0303	0.0266 (28559)	0.0347
Jun-05	-0.0293	-0.0295 (60560)	0.0613
Jul-05	0.0353	0.0277 (31758)	0.0275
Aug-05	0.0406	0.0284 (33740)	-0.0283
Sep-05	0.0326	0.0219(39964)	-0.0280
Oct-05	-0.0013	0.0248(47982)	-0.0615
Nov-05	-0.00045	0.0391(74970)	0.0618
Mean return	0.0123	0.0175	0.0022
Standard deviation	0.0276	0.0299	0.0530
Correlation with market	-0.209	-0.097	
Pearson P-Value	0.316	0.646	

Table 2. Return rates of Model-1 and Model-2 with MTX

Month	Model-1/MTX	Model-2/MTX (target t)	market
Nov-03	-0.0009	0.0090 (18173)	-0.0099
Dec-03	0.0118	0.0129 (17923)	-0.0194
Jan-04	0.0134	0.0543 (18095)	0.1100
Feb-04	0.0449	0.0034 (18919)	0.0346
Mar-04	0.0158	0.0121 (12781)	-0.0042
Apr-04	0.0134	0.0290 (7543)	0.0353
May-04	0.0390	0.0374 (9155)	-0.1394
Jun-04	0.0350	0.0344 (6175)	-0.0513
Jul-04	0.0350	0.0346 (8596)	-0.0272
Aug-04	0.0352	0.0304 (15484)	0.0034
Sep-04	0.0436	0.0569 (16906)	0.0817
Oct-04	0.0122	0.0567 (16346)	-0.0141
Nov-04	-0.0398	-0.0284 (9475)	0.0415
Dec-04	0.0053	0.0027 (10114)	-0.0043
Jan-05	0.0344	0.0316 (12071)	-0.0179
Feb-05	-0.0425	-0.0425 (11652)	0.0421
Mar-05	-0.0113	0.0088 (8183)	-0.0116
Apr-05	-0.0447	-0.0247 (7897)	-0.0625
May-05	-0.0112	0.0994 (13312)	0.0347
Jun-05	-0.0253	-0.0380 (7579)	0.0613
Jul-05	0.0447	0.0247 (5105)	0.0275
Aug-05	0.0370	0.0337 (8193)	-0.0283
Sep-05	0.0325	-0.0011 (7523)	-0.0280
Oct-05	-0.0048	0.0186 (7659)	-0.0615
Nov-05	0.03137	0.0539 (21563)	0.0618
Mean return	0.0122	0.0204	0.0022
Standard deviation	0.0283	0.0326	0.0530
Correlation with market	-0.103	0.063	
Pearson P-Value	0.624	0.765	

returns of the assets in the coming month in the second stage are then simulated by using single index model. Single index model describes the relation of the return of each asset with the market in a simple linear regression model. We only need to simulate the return of the market as a Brownian motion and the corresponding error term in each regression model. We call the model with fixed target Model-1 and the one with dynamic target Model-2. In each model, we further consider one of the two index futures(TX or MTX). Therefore, four models, Model-1 with TX, Model-2 with TX, Model-1 with MTX and Model-2 with MTX, are tested by using MPL/Cplex software. We assume our budget is NT\$1,200,000(around US\$40,000) for the model with TX and NT\$ 300,000 (around US\$ 10,000) with MTX. For Model-1, the corresponding fixed targets for TX model and MTX model are NT\$36,000 and NT\$9,000, respectively. We let C be zero and α be 0.15.

Our SLP models have been tested for each of the consecutive 25 months starting from Nov. 2003. During each month, we may close the portfolio at any day once the criterion is satisfied. The numerical performance results of Model-1 and Model-2 with TX and MTX are presented in Table 1 and Table 2, respectively, and the corresponding graphical comparisons diagrams are shown in Figure 1 and Figure 2.

The mean monthly return rates of Model-1 with TX, Model-1 with MTX, Model-2 with TX and Model-2 with MTX are 1.23%, 1.122%, 1.75% and 2.04%,

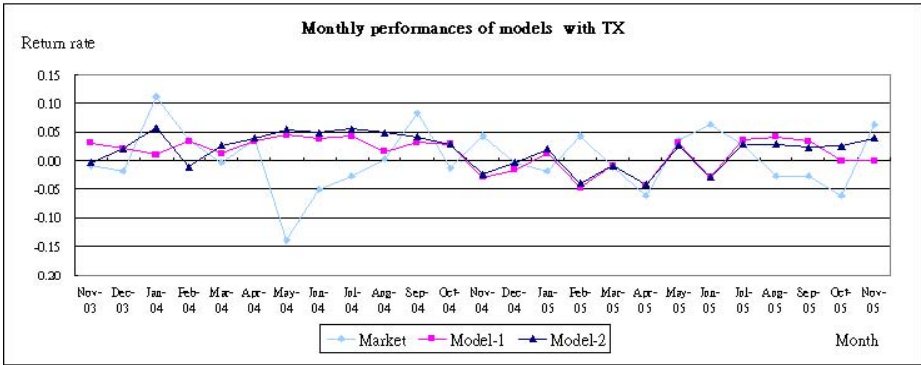


Fig. 1. Comparisons of the return rates between Model-1/TX, Model-2/TX and market

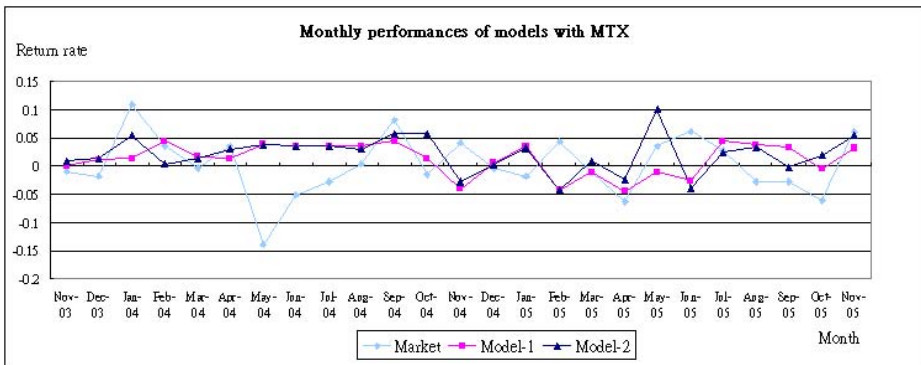


Fig. 2. Comparisons of the return rates between Model-1/MTX, Model-2/MTX and market

respectively. They are all far higher than the local monthly interest rate of savings, 0.1071%. Although beating the market is not what we concern, the results show that our returns are also better than the market return, 0.22%.

The standard deviation of the return rates of Model-1 with TX, Model-1 with MTX, Model-2 with TX and Model-2 with MTX are 2.76%, 2.76%, 2.99% and 2.69, respectively. They are all significantly smaller than the standard deviation of the return rate of market, 5.33%. Also, by observing Figure 1 and Figure 2, the monthly returns of our portfolios fluctuates in a smaller scale than the market does. It indicates that the returns of our portfolios are steadier than market, furthermore, the correlations between each of these four models and market are all small and, according to Pearson test, the large P-values indicate that the return rates from our models are uncorrelated with the market.

4 Conclusion

This research provides two-stage stochastic program models for selecting the portfolio along with one index future. These models are also incorporated with other statistical models such as single index model and MAD model. Empirical results show that the returns of the portfolio under our investment strategy are stable, independent of the market and higher than the interest rate of savings. These are the important investment criterions for risk-aversion investors. Two-stage SLP shows itself a sophisticated way for solving the portfolio selection problem and it works well. Retaining the framework of two-stage SLP, multiple index model and MAD model for the skewness can further be considered in the future study.

References

1. Cai, X., Teo, K. L., Yang, X. and Zhou, X. Y. (2000). Portfolio Optimization under A Minimax Rule, *Management Science*, Vol. 46, No. 7, 957-972.
2. Chang, K.-H., Chen, H.-J. and Liu, C.-Y. (2002). A Stochastic Programming Model For Portfolio Selection, *Journal of the Chinese Institute of Industrial Engineers*, Vol. 19, No. 3, 31-41.
3. Chang, K-H. (2004). Safety-First Portfolio Selection Problem with Index Future, Technical report, Dept. of Industrial Engineering, Chung Yuan Christian University.
4. Elton, E. J. and Gruber, M. J. (1995). *Modern Portfolio Theory and Investment Analysis*, John Wiley and sons, New York.
5. Higle, J. L. and Sen, S. (1996). *Stochastic Decomposition*, Kluwer Academic Publishers, Dordrecht
6. Kane, A. (1982). Skewness Preference and Portfolio Choice, *Journal of Financial and Quantitative Analysis*, Vol. 17, 15-25.
7. Konno, H. and Yamazaki, H. (1991). Mean-Absolute Deviation Portfolio Optimization Model and Its Applications to Tokyo Stock Market, *Management science*, Vol. 37, No. 5, 519-531.
8. Konno, H. and Kobayashi, K. (1997). An Integrated Stock-Bond Portfolio Optimization Model, *Journal of Economic Dynamics and Control*, Vol.21, 1427-1444
9. Konno, H, Shirakawa, H. and Yamazaki, H. (1993). A Mean-absolute Deviation-Skewness Portfolio Optimization Model, *Annals of Operations Research*, Vol. 45, 205-220.
10. Kouwenberg, R. (2001). "Scenario Generation and Stochastic Programming Models for Asset Liability Management, *European Journal of Operational Research*, Vol. 134, 279-292.
11. Markowitz, H. M. (1952). Portfolio Selection, *Journal of Finance*, Vol. 7, 77-91
12. Telser, L. G. (1955). Safety First and Hedging, *Review of Economics Studies*, Vol. 23, 1-16.

Hierarchical Clustering Algorithm Based on Mobility in Mobile Ad Hoc Networks

Sulyun Sung, Yuhwa Seo, and Yongtae Shin

Dept. of Computer Science, Soongsil University, Sangdo-Dong, Dongjak-Gu,
Seoul 156-764, Korea
{ssl, zzarara, shin}@cherry.ssu.ac.kr

Abstract. This paper proposes a hierarchical clustering method based on the relative mobility pattern in mobile ad hoc environment. The relative mobility pattern between two nodes is evaluated by using a received message and the cluster is created by grouping nodes with a relative mobility below a specific threshold. Also, we create a hierarchical clustering structure by allowing a merge among clusters based on the mobility pattern. In this way, the proposed mechanism can increase a continuity of total cluster structure. Since we allow the combination of clusters, we can reduce the number of cluster and message required for a routing. To evaluate a performance of our mechanism, we compared ours with the existing LCC and WCA by a Glomosim. The simulation results show that our scheme can provide the higher stability and efficiency than existing schemes.

1 Introduction

Ad hoc network is a dynamically reconfigurable wireless network with no fixed infrastructure or central administration. Each host acts as a router and supports a network function such as a traffic routing. For two nodes to communicate in ad hoc network, the routing on the wireless path of multi-hop is required. But, since ad-hoc environment doesn't have a fixed infrastructure and the wireless link between two nodes cannot be moved, many problems may happen. The node mobility results to a frequent communication disconnection and the path re-setup by a node movement causes the network congestion. The efficiency of an adaptive routing algorithm depends on a timeliness and topology information. But the most important factor is the number of information exchange. Since the ad hoc environment has a frequent topology change, the exchange of updated routing information can make the network be saturated easily. Since the rate of link failure depends on a mobility of node, the high mobility will increase a traffic consumed to maintain a path. Therefore, the solution for an efficient routing is to minimize a response to mobility. To provide multi-hop communication in ad hoc networks, numerous routing protocols have been developed. Many of these protocols can be placed into one of two classes: proactive (table-driven) approaches, and reactive (on-demand) approaches. Proactive protocols are derived from the traditional distance vector and link state protocols commonly used in wired networks. These protocols maintain routes between each pair of nodes throughout the lifetime of the network. While this approach has the benefit that a

route is generally available the moment it is needed, proactive protocols have poor scaling properties due to their $O(n^2)$ overhead. Additionally, previous work has shown these protocols to not perform as well as reactive routing protocols in most scenarios. Reactive protocols, on the other hand, only establish routes on-demand, or when needed. These protocols thereby only incur overhead for route construction and maintenance when those routes are actually needed, since they do not maintain routes that are not utilized. The drawback to these protocols is that they introduce a route acquisition latency, or a period of waiting to acquire a route after the route is needed. These protocols have shown to also have limited scalability, due to their route discovery and maintenance procedures. One alternative to these protocols for improving scalability is clustering, or hierarchical, routing protocols. Hierarchical protocols place nodes into groups, often called clusters. These groups may have some sort of cluster leader that is responsible for route maintenance within its cluster and between other clusters. Also, the effect of dynamic topology change is limited to the inside of cluster. The reactive routing is used within the cluster and the proactive routing to deliver a data between clusters is used. The most important thing in these routing protocols is to reduce an overhead to create and maintain a cluster. If the clustering algorithm is very complex and cannot guarantee a stability of cluster, it cannot cause an efficient result. Also, when the mobility is very high, the overhead by a cluster re-generation will increase. Therefore, the mobility of node should be considered importantly when the cluster is created.

This paper proposes a hierarchical clustering method based on the relative mobility pattern of the neighboring nodes in mobile ad hoc environment. The relative mobility pattern between two nodes is evaluated by using a received message and the cluster is created by grouping nodes having the mobility pattern below a specific threshold. Also, we create a hierarchical clustering structure by allowing a merge among clusters based on the mobility pattern. In this way, the proposed mechanism can increase a continuity of cluster and since we allow the combination of clusters, we can decrease the number of cluster and message required for a routing. The rest of the paper is organized as follows. Section 2 presents a new hierarchical clustering method based on the mobility pattern. Section 3 demonstrates the performance improvement of our scheme over the existing scheme using glomosim. Finally, we conclude in Section 4.

2 Clustering Algorithm Based on Mobility Pattern

The proposed clustering algorithm uses a mobility pattern of mobile node as a major metric for a cluster formation. The proposed algorithm to improve a durability of created cluster uses relative mobility with neighboring nodes to be worked based on mobility pattern of each node. We can define that two nodes that have the similar velocity and direction relatively have the less relative mobility difference. One cluster is formed by nodes which have the less relative mobility difference and the more adjacent clusters than two can be combined if they have a similar relative mobility difference. As a result, since the mobile nodes with a similar mobility are grouped, the formed cluster in our scheme can be kept continuously more than one in existent clustering techniques. Also, as nodes incline to move by a unit of group in ad-hoc network that support multiplex connection communication, the proposed clustering

algorithm can be applied more effectively[6],[7]. Because clusters with the less relative mobility difference can be combined, the proposed algorithm can reduce the number of cluster in all ad-hoc networks efficiently. This paper supposes following facts. Two nodes are connected by full-duplex link and each node can measure the signal strength. Also, the network topology can be defined by graph $G=(V,E)$. At this time, V is set of node, and E is set of full-duplex links that act independently in each direction. The logical distance $d(x,y)$ between two nodes belonging to graph G is defined by the number of minimum hop and becomes $d(x,y) \leq L$ of two nodes selected randomly in one cluster. Also, the real distance between adjacent two nodes x,y can be measured by a received signal strength by a hello message or beacon message that exchange periodically. The received electric power can be calculated by numerical formula (1) by using a Friss's freedom space damage formula.

$$P_r = P_t \times G_t \times G_r \times \frac{\lambda^2}{(4 \times \pi \times d)^2} \quad (1)$$

P_r in (1) is received electric power, P_t is transmitted electric power, G_t is electric power gains of sending antenna (dB), G_r is electric power gains of receiving antenna (dB), λ is utilization wave length, d is distance. Generally, since an electric power is proportional in square of distance, the real distance between two nodes can be calculated by P_r of (1). As it is impossible to calculate correct distance by using the received electric power, the produced distance becomes a distance estimate.

2.1 Relative Mobility Calculation

The proposed algorithm utilizes mobility model of [5] to calculate a relative mobility. The most general mobility model is Random Walk Mobility Model. We basically assume that the mobile node follows a Random Walk Mobility Model to increase a generality. Each mobile node identifies existence of neighboring nodes through hello message. When the mobile node received hello message, it knows the only location information at time t . Therefore, to measure relative mobility with neighboring nodes, the mobile node uses location information collected for specific time.

If the movement pattern of neighboring nodes becomes known by other additional method, each mobile node can calculate a relative mobility by an estimate vector defined in [5] without collecting a location information for a specific time. First, this paper supposes that it is impossible to predict a next location of each node. Therefore, this section describes a method to calculate relative mobility by Random Walk Mobility Model. Since the mobile node which follows a Random walk mobility model selects a direction and speed randomly, ad hoc node cannot predict a direction vector after t time.

Each node selects the new speed and direction in predefined extent [the minimum speed, the maximum speed] and $[0, 2\pi]$. The mobile node exchanges a hello message periodically with neighborhood nodes. The mobile node can calculate the received electric power using (1) from hello message. The node A can calculate distance $E[d_{AB}]_t$ at time t like a formular (2) by a received electric power.

$$E[D_{AB}]_t = \frac{k}{\sqrt{P_r}} = \frac{k}{\sqrt{P_t \times G_t \times G_r \times \frac{\lambda^2}{(4 \times \pi \times d)^2}}} \quad (2)$$

The relative mobility RM_{AB} between A and B shows a distance difference about whether A and B moved relatively in some degree. When RM_{AB} value is smaller, we can conclude that two nodes A, B are moving to more similar direction with similar speed. If a distance difference between node A and B during t_1-t_0 is calculated, the relative mobility of two nodes for t_1-t_0 , RM_{AB} can be produced.

$$RM_{AB} = E[D_{AB}]_t - E[D_{AB}]_{t-1} \quad (3)$$

Also, the average relative mobility, RM_{AB-T} during a specific time T between node A and B can be calculated like a formula (4). Each node updates neighboring node management table by using a (4).

$$RM_{AB-T} = \frac{1}{N} \sum_{i=1}^N (E[D_{AB}]_i - E[D_{AB}]_{i-1}) \quad (4)$$

2.2 Clustering Algorithm

Each node transmits a hello message periodically. Each node can identify the existence of neighboring node by a received hello message and decide a role of itself in cluster. The Hello message each node transmits includes <node ID, Mode, Cluster Head ID, Seq_num, Level_info>. At this time, node ID is unique identification data that each node is given, Mode has one value among 4 values(cluster head, general node, gateway, relay gateway). If one node belongs in more cluster head than two, the mode of this node becomes gateway.

The relay gateways are nodes that lie in separate clusters, but that are within transmission range of one another. Such pairs of nodes can also be used to route between clusters. Cluster Head ID is ID of cluster head node of itself and seq_num is used to identify a hello message and is augmented by 1 whenever is transmitted. Level_info is a value for organizing clusters hierarchically and will be explained in detail in 3.2.1 chapter.

Node can not decide own mode when initialized. Therefore, each node sets up its mode to a general node and informs own existence to neighboring nodes by a hello message. After each node transmits hello message, it sets up a timer to receive a hello message from other node. After each node receives a hello message from neighboring node for a specific time, it creates neighboring node management table by this message. The information in a neighborhood node management table is ID of node that transmits hello message, electric power P_r produced by (1), relative mobility at time t produced by (3), relative mobility values during limited time T produced by (4). After each node receives hello message, it measures the signal strength of received message and creates new entry in neighborhood node management table by using node id in message. After each node adds the signal strength in neighborhood node management table, it stores a distance produced by (2) and time t together in neighborhood node management table. If the neighborhood node management table

already has the id of correspondent node, each node produces a relative mobility by (3) and a distance value of time $t-I$ and stores this in neighborhood management table.

Each node runs this procedure repeatedly until the predefined timer is expired. Finally, when timer is expired, the mobility node can get an average relative mobility with neighboring node for time T. If the set of node that exists on neighboring node management table is S_m , each node selects node B which has the smallest $RM_{AB,T}$, $RM_{AB,T} < Threshold_{mob}$, $B \in S_m$ and sends a Head_request message to node B. In other words, the cluster head is decided by (5). At this time, the mobility threshold value, $Threshold_{mob}$ is a design parameter of algorithm and is determined by an experiment since it is a variable to control a stability of cluster.

$$Cluster_Head = Least_{i \in S_m} \{ ID \mid RM_{AB,T} < Threshold_{mob} \} \tag{5}$$

The node that receives hello message updates a neighboring node table first. If the mode in received hello message is cluster_head and the cluster_head_id of itself is undefined, it validates whether $RM_{AB,T} < Threshold_{mob}$. If $RM_{AB,T} < Threshold_{mob}$, it transmits cluster_ack message to the node which transmits a hello message and updates own neighboring node table. But if the mode of hello message received is not cluster_head or own cluster_head_id is not set up yet, the node stores a hello message in cache and updates a neighboring node management table after calculating a mobility. Each node repeats this work during T time. If RM value satisfies $Least_{i \in S_m} \{ ID \mid RM_{AB,T} < Threshold_{mob} \}$ after T time, each node sends head_request message to the relevant node. After node that receives a head_request message corrects own Mode to cluster head, it transmits hello message to nodes in the neighborhood node table and cluster member table. At this time, the hello message is transmitted with a short interval to reduce a time required to form a cluster. If the node does not find a node that has smaller $RM_{AB,T}$ than $Threshold_{mob}$ value during a limited time, it set up itself to a cluster head and transmits a hello message. Also, the node that receive hello message from the more cluster heads than two sets up its mode to gateway.

Algorithm. 1-level cluster formation

Variables

- Neighboring_node_list_i = Node I's neighboring node management table
- Cluster_member_list_i = Cluster head, node I's cluster member table
- $RM_{ij,T}$ = Relative mobility between node i and node j
- node_h = Node h that has the smallest $RM_{ih,T}$ value than $Threshold_{mob}$ among Neighboring_node_list_i of node i

Initially Neighboring_node_list_i = nil, Cluster_member_list_i = nil
upon timer is expired;

Send Hello message to \forall nodes \in Neighboring_node_list_i
upon receiving Hello message from node_j ;
update Neighboring_node_list_i ;
if(hello_j->mode = cluster_head && cluster_head_id_i = undefined)
 if($RM_{ij,T} < Threshold_{mob}$)
 Send cluster_ack message to node_j
 Update Neighboring_node_list_i
else

```

while(timer of  $T$  is not expired)
  Save Hello message in cache
  Measure power signal and calculate and save  $RM_{ij-T}$ 
  If( $\text{node}_j == \{Least_i \in S_m \{ID \mid RM_{AB-T} < Threshold_{mob}\}\}$ )
    Send Head_request message to  $\text{node}_j$ 
upon receiving cluster_ack message from  $\text{node}_j$ 
  update cluster_member_list $_i$ 
upon receiving Head_request message from  $\text{node}_j$ 
  update  $\text{mode}_i$  to cluster_head
  update Cluster_member_list $_i$ 
  Send Hello message to  $\forall \text{nodes} \in \text{Neighboring\_node\_list}_i \&\& \text{Cluster\_member\_list}_i$ 

```

2.2.1 Hierarchic Cluster Formation

The Hello message that each node transmits includes Level_info value. This level_info value is increased by 1. The nodes that are not assigned a Level_info have undef value. All nodes except a cluster head have a Level_info value of 0. The cluster head of Level-1 has Level_info value of 1, and cluster head of Level-2 has Level_info value of 2. The value of Level_info is same with the hierarchical level of relevant cluster head.

The cluster gateway and relay gateway receives a hello message periodically from node in neighboring clusters. Whenever both a gateway and a pair of joint gateways exist to connect two clusters, the single gateway is favored because it is one fewer hop. Since each hello message includes cluster head ID and Level_info value, each node can hold the hierarchical level of each cluster.

After the gateway receives hello message from adjacent nodes, it creates entry on neighboring node management table. At this time, the gateway and relay gateway may have the entry of node in other cluster. If the relative mobility with node in other cluster is smaller than $Threshold_{mob}$, two adjacent clusters can organize a hierarchical structure. By a hello message, the node that has smaller cluster head ID transmits Cluster_Join message to neighboring gateway and own cluster head to form a hierarchical structure. At this time, the relay gateway transmits Cluster_Join message to neighboring relay gateway and cluster head. The Cluster_Join message includes the hierarchical information of cluster of a sending

The node that receives Cluster_Join message forwards a Cluster_Join message to cluster head. After receiving a Cluster_Join message, a cluster head sends a hello message after increasing its Level_info by 1. Basically, before forming a hierarchical structure, all cluster heads have Level_info value of 1. The node that has Level_info value of 1 becomes always a parent cluster and the cluster that has Level_info value more than 2 becomes child cluster of parent node. In this way, the hierarchical clustering structure is formed. In case that two cluster heads are located within a direct transmission range, they also can form a hierarchical structure. Though one of two cluster heads gives up a head and two clusters can be combined, this causes a ripple effect to whole ad hoc network. Therefore, this paper supposes that if there is two cluster heads in direct transmission range and the relative mobility is lower than $Threshold_{mob}$, the hierarchical clustering mechanism is used. At this time, in case that two cluster heads have same Level_info value, they decide a hierarchy using node id. And, in case that two cluster heads have a different Level_info value, the node that

has big id adds 1 to Level_info of node id which has a small id relatively and set up the result value to its Level_info. After that, the node that has big id notifies a Level_info change by sending a hello message.

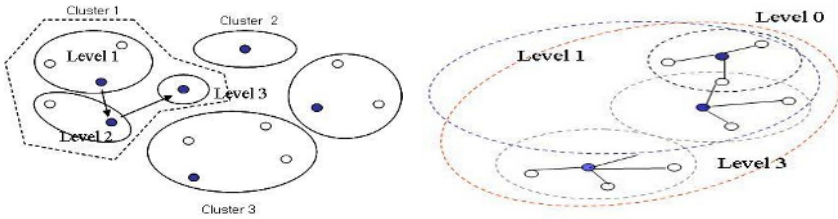


Fig. 1. Hierarchical cluster formation

After CH1 and CH2 move as an example (c) of figure 2, if they belong within each other's transmission range, CH2 that has bigger id transmits Cluster_head_join message to CH1 and set up own Level_info to Level_info of CH1 + 1. CH2 transmits hello message to nodes in a neighboring node management table to notify a Level_info change. But, in case that CH1's cluster includes CH2's cluster as an example (4) of figure 2, CH2 gives up mode of cluster head and returns to a general node.

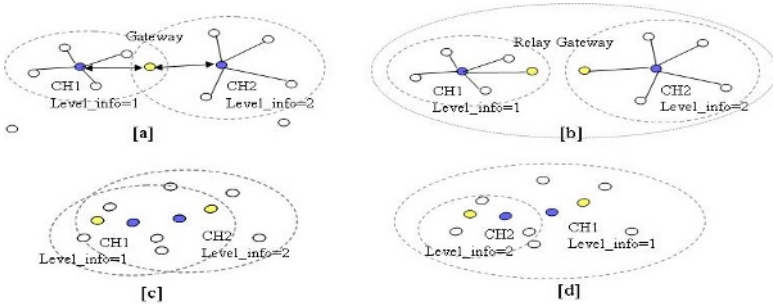


Fig. 2. Hierarchical clustering structure

Algorithm. Hierarchical cluster formation

```

Initially Neighboring_node_listi != nil, Cluster_member_listi != nil
upon receiving Hello message from nodej ;
  if(helloj->cluster_head_id!=cluster_head_idi && helloj->mode!=cluster_head)
    while(timer of T is not expired)
      Save Hello message in cache
      Measure power signal and calculate and save RMij,t
      if(RMij,t < Thresholdmob && node_idi < node_idj)
        Send cluster_join message to nodej
      else if(helloj->cluster_head_id!=cluster_head_idi && helloj-> mode=cluster_head)
  
```

```

if(Cluster_member_listi ⊆ Cluster_member_listj)
  while(timer of T is not expired)
    Save Hello message in cache
    Measure power signal and calculate and save  $RM_{ij,t}$ 
    if( $RM_{ij,t} < Threshold_{mob}$  && node_idi < node_idj)
      Send cluster_head_join message to nodej
upon receiving cluster_join message from nodej
  if (modei = gateway || modei = joint_gateway)
    Forward cluster_join message to cluster head
  else if (modei = cluster_head)
    set to Level_infoi = level_infoi + 1;
    Send Hello message to  $\forall$  nodes  $\in$  Neighboring_node_listi &&
      Cluster_member_listi
upon receiving cluster_head_join message from nodej
  if(node_idi > node_idj)
    set to level_infoi = level_infoj + 1;
    send Hello message to  $\forall$  nodes  $\in$  Neighboring_node_listi &&
      Cluster_member_listi

```

3 Performance Evaluation

Each of the experiments was performed using the Glomosim network simulator[9] developed at UCLA. The IEEE 802.11 MAC layer protocol is used for channel access in each simulation. For the performance evaluation of proposed mechanism, we compares ours with LCC and WCA by Jorge Nuevo[11]. Glomosim is developed at the University of California, Los Angeles using PARSEC[10]. The simulation is executed for 500 seconds. The node density is fixed to 80 and transmission range of each node is limited by 350. The mobility model is set to a random direction model. The maximum speed of a node is varied between 0m/s and 10m/s. The $Threshold_{mob}$ is set to $M_{mob} + k \cdot \delta_{mob}$ ($k=1.5$). The simulation is executed to evaluate a stability and efficiency. We measured the number of cluster change according to a speed of a node to evaluate a stability of our mechanism and the control message number according to a speed to evaluate an efficiency of our mechanism. To improve the stability of our mechanism, we compared ours with existing schemes, LCC, WCA. [Figure 3] shows a comparison result of cluster change. In this simulation, the number of node is fixed to 50. Since the cluster change is resulted from a cluster head change, we measured the nodes changed from the general node to the cluster head and the nodes changed reversely. As the average node speed increases, the number of such changes rises. At higher speeds, nodes change neighbors more rapidly. The LCC results in the greatest number of leadership changes during the simulation. Our mechanism results in the least number of cluster changes. In this result, we can show that our mechanism can provides the higher stability than existing schemes. The stability of the cluster topologies is more closely reflected by the second one [figure 3]. This figure represents the number of cluster leaders at each second during the first 170 seconds of

the simulations. Our mechanism results in a fairly stable number of cluster leaders that fluctuates between 11 and 14. But, LCC and WCA show much wider variance. A high number of cluster leader changes results in significant communication overhead for routing updates and cluster reconfigurations, and leads to an unstable topology.

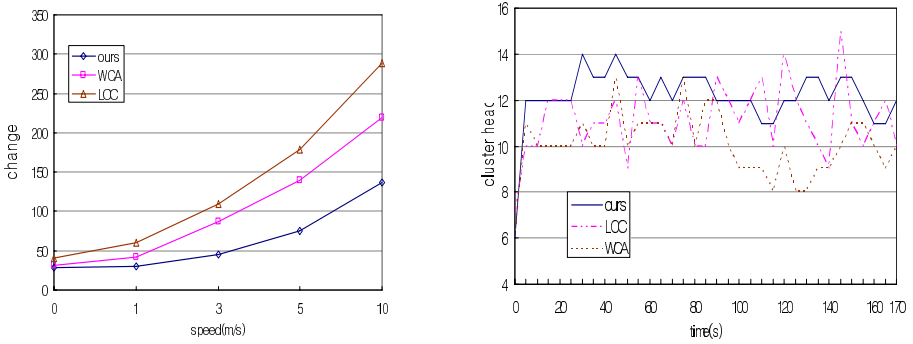


Fig. 3. Number of cluster and cluster head

[Figure 4] shows a simulation result of efficiency of our mechanism. The first one of [Figure 4] represents the number of packets able to be delivered by our scheme running over AODV. AODV-C means the AODV with our clustering scheme. In figure 4, our scheme outperforms AODV without our clustering scheme. Since our clustering scheme can provide a stable backbone, the speed of node have a little effect on our scheme relatively. The second one of [Figure 4] shows the number of RREQ.

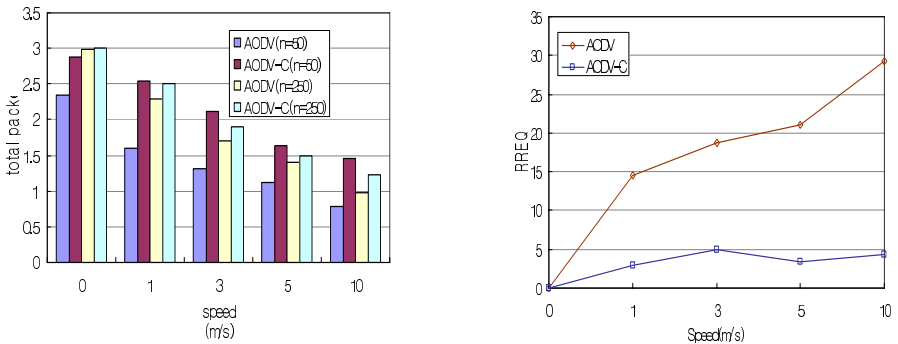


Fig. 4. Total data packet and RREQ comparison result

In our scheme, the all control messages are processed by a cluster head. Therefore, our scheme has far fewer control messages than AODV without our scheme.

4 Conclusion

This paper proposes a hierarchical clustering method based on the relative mobility pattern of the neighboring nodes in mobile ad hoc environment. The relative mobility pattern between two nodes is evaluated by using a received message and the cluster is created by grouping nodes having the mobility pattern below a specific threshold. Also, we create a hierarchical clustering structure by allowing a merge among clusters based on the mobility pattern. In this way, the proposed mechanism can increase a continuity of cluster. Since we allow the combination of clusters, we can reduce the number of cluster and message required for a routing. To evaluate a performance of our mechanism, we compared ours with the existing LCC and WCA[10] by a Glomosim. The simulation results show that our scheme can provide the higher stability and efficiency than existing schemes.

References

- [1] S.Basagni, Distributed clustering for ad hoc networks, in:Proceedings of the 1999 International Symposium on Parallel Architectures, Algorithms, and Networks, Australia (June 1999) pp. 310-315
- [2] E.M Belding-Royer, Hierarchical routing in ad hoc mobile networks, to appear in the Wireless Communications and Mobile Computing (2002).
- [3] C.-C. Chiang, H.-k, Wu, W. Liu and M.Gerla, Routing in clustered multihop, mobile wireless networks with fading channel, in: Proceedings of IEEE Singapore International Conference on Networks(SICON) (April 1997) pp. 197-211
- [4] X. Hong, M.Gerla, G. Pei, and C. Chiang. A group mobility model for ad hoc wireless networks, In Proceedings of ACM/IEEE MSWiM, Seattle, WA, Aug. 1999
- [5] R. Ramanathan and M.Steenstrup, Hierarchically-organized, multihop mobile wireless networks for Quality-of-service support, ACM/Baltzer Mobile Networking and Applications 3(1) (1998) 101-118
- [6] Hass Zj. A new routing protocol for the reconfigurable wireless networks. Proceedings of the ICUPC '97 1997
- [7] Lee S-J, Su W. Hsu j, Gerla M, Bagrodia R. A performance comparison study of ad hoc wireless multicast protocols. Proceedings of the IEEE INFOCOM 2000
- [8] Boris Mitelman and Arkady Zaslavsky. Link State Routing Protocol with Cluster Based Flooding for Mobile Ad-hoc Computer Networks. In Proceedings of the workshop on Computer Science and Information Technologies(CSIT), Moscow Russia, 1999
- [9] GloMoSim: Global Mobile Information Systems Simulation, <http://pcl.cs.ucla.edu/projects/glomosim/>
- [10] PARSEC: Parallel Simulation Environment for Complex Systems, <http://pcl.cs.ucla.edu/projects/parsec/>
- [11] C.R.Lin and M.Gerla. Adaptive clustering for mobile wireless networks. IEEE Journal on Selected Areas in Communications, 15(7):126501275, Sept. 1997
- [12] A.D.Amis, R. Prakash, T.H.P. Vuong, Max-min d-cluster formation in wireless ad hoc networks. In proceedings of IEEE INFOCOME '00, Vol. 1, pages 32-41, Mar. 2000

An Alternative Approach to the Standard Enterprise Resource Planning Life Cycle: Enterprise Reference Metamodeling

Miguel Gutiérrez¹, Alfonso Durán¹, and Pedro Cocho²

¹ Departamento de Ingeniería Mecánica, Universidad Carlos III de Madrid,
Av. de la Universidad 30, 28911 Leganés (Madrid), Spain
{miguel.gutierrez, alfonso.duran}@uc3m.es
<http://www.uc3m.es/uc3m/dpto/dpcleg1.html>

² Adalid MyO,
c/Berlín 3F, of. 1.16, 30395 Cartagena (Murcia), Spain
pcocho@adalidmyo.com
<http://www.adalidmyo.com>

Abstract. The Enterprise Resource Planning (ERP) systems development is based on the initial definition of an enterprise reference model, whose richness and generality will basically determine the flexibility of the resulting software package to accommodate specific requirements. The desired accommodation is rarely accomplished without costly customization processes, which are frequently unaffordable for the small and medium enterprises. In this paper, an alternative ERP development approach, stemming from the implementation of an enterprise reference metamodel, is proposed. We present an analysis of the resulting alternative ERP life cycle, particularly focusing on the metamodel that constitutes the core of the proposal. The results obtained in a complete enterprise software development project, encompassing software package development, tailored implementation and some post-implementation customization, show that the proposed approach facilitates the identification and fulfillment of the customer's specific requests at a reduced cost, even if they arise after the implementation.

1 Introduction

Enterprise Resource Planning (ERP) systems are commercial software packages that provide an integrated solution to support all the areas of a company (financial, human resources, production...) [1]. The solution starts from a generic reference model of the business entities of a company and their functional relationships [2]. A well known example is the set of models proposed by Scheer [3]. Flexibility at the implementation stage is basically contingent on the richness of the initial model [4]. Where the accommodation to the enterprise requirements is not satisfactory, as it is normally the case, a code customization process is required [5], [6]. This is particularly onerous to the small and medium enterprises (SMEs) that frequently cannot afford it [5], [7].

Even when that customization is not required, after the software has been in production for some time new requirements will eventually appear. The post-implementation flexibility of ERP systems, that is, the ability to accommodate these

new requirements, is, however, very limited. In execution time, no new entities can be added to the initial model, no new attributes assigned to the existing entities, and no new functional relationships established.

This paper is an outcome of a research project undertaken to tackle this lack of flexibility exhibited by enterprise software by allowing, at execution time, not only the creation of new instances of previously modeled entities, but also the definition of new entities, the definition and assignment of new attributes, and the creation of new relationships. This approach leads to increasing by one layer the abstraction level of the enterprise model, i.e., shifting the codification boundary to the metamodel layer [8], which, in turn, leads to an alternative ERP life cycle.

In the following section, a schematic representation of the basic phases encompassed by the prevailing ERP life cycle is depicted as a reference, and the current related research literature is briefly reviewed. In the third section, the metamodel that constitutes the core of the proposal is described, and it is positioned in the four-layer metamodel architecture. The fourth section presents the proposed alternative software life cycle based on the abovementioned metamodel, highlighting its conceptual and practical differences. The fifth section presents the experimentation carried out in a real development project, to finalize in the conclusions section by summarizing the major advantages and implications of the proposed approach.

2 Enterprise Software Systems Development Cycle. Current Status

ERP systems do not adhere to a fixed life cycle. However, a generic sequence of stages can be identified (Fig. 1) [1], [4], [5], [6], [7]:

- **Reference enterprise model implementation.** The starting point is a reference model of an enterprise—a superset of the models of the spectrum of target organizations—depicted in Figure 1 as a network of interrelated elements.
- **Industry-specific solution.** In order to facilitate implementations, the major ERP vendors have adopted two strategies: a modular design which allows building a first cut of a specific solution by combining some basic modules (like the Solution Composer of the ERP market leader SAP [9]), and the development of some pre-packaged industry-specific solutions, such as pharmaceutical, chemical, automotive or insurance [10], [11]... often used as the implementation starting point.
- **Company-specific submodel – Parametrization.** Whether the starting point is the generic or the industry-specific model, the next implementation stage involves distilling the submodel that corresponds to the specific requirements of the company that is implementing the system. This normally involves a substantial and costly consulting engagement, which is aimed at establishing the set of parameters (more than 5000 in SAP R/3 [7]) that particularize the initial model.
- **Code Customization.** Even after parametrization, the model normally does not address all the specific requirements, thus forcing the company to either modify/supplement the software or/and change its business processes and procedures to conform to those embedded in the software. This requires expensive ad-hoc devel-

opment, depicted in Figure 1 through the addition of new elements and the modification (represented through a color/pattern change) of existing ones.

- **Post-implementation.** Some companies embark in successive ad-hoc developments in the post-implementation stage, also involving substantial outlays. This is depicted in Figure 1 through an additional color/pattern change in one element.

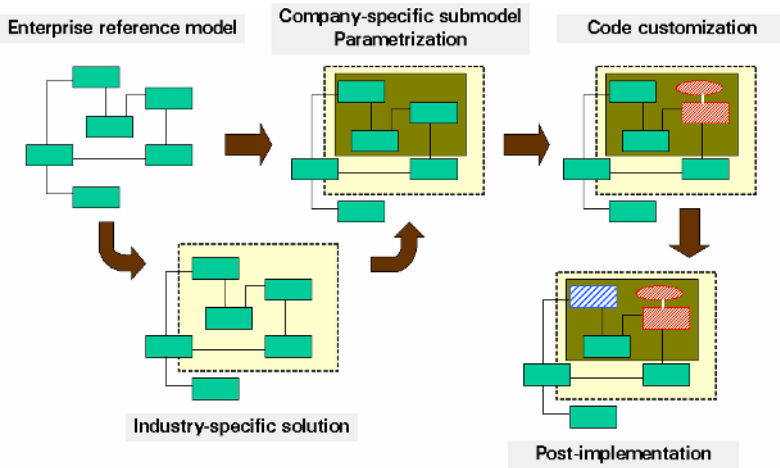


Fig. 1. Generic ERP life cycle

These stages encompass the basic foundations of the current ERP life cycle. Further refinements proposed by some authors provide some useful insights. Brehm *et al.* propose a typology of what they call ERP tailoring types, including the role played by third-party packages (bolt-ons) and legacy systems [4]. Luo and Strong combine the standard stages with three options in process customization (no change, incremental, radical change) [5]. Scheer and Habermann highlight the problem of implementation costs, stating that the major ERP vendors estimate those costs (processes alignment plus software customization) in three to seven times the cost of the software license [7]. In summary, what stands out from the literature review is the troublesome nature of ERP implementations, the criticality of the alignment between the systems and the enterprise processes [6], and the generalized need to carry out customization processes to fill the gap between the required model and the ERP reference model.

To ease the implementation process, the current efforts in ERP systems are primarily directed towards improving the enterprise modeling languages. Soffer *et al.* analyze the desirable characteristics of such languages [12]. Rosemann and van der Aalst discuss the limitations of the current enterprise modeling languages and propose an improved version of the Event-Driven Process Chains (EPCs) used by SAP [2]. A valuable reference is the research carried out by the UEML (Unified Enterprise Modeling Language) group. UEML combines both the intention of standardizing the enterprise modeling languages and of providing a framework to information sharing between enterprises [13]. The latter is aligned with the current research focus on applying metamodeling to enterprise software, which is providing a common repository of information to be shared by enterprises or for data warehousing purposes [14].

3 Enterprise Metamodeling

This paper takes an approach that differs from the abovementioned: to base the software development in the implementation of an enterprise reference metamodel. To facilitate the positioning of the proposed approach, we will first describe the conventional four-layer metamodeling architecture and then the enterprise metamodel, including an example of application in the reference architecture.

3.1 Four-Layer Metamodeling Architecture

The conventional four-layer metamodeling architecture establishes the basic metamodeling concepts through a hierarchy of modeling levels or (meta)layers (M_n ; $n=0,1,2,3$), in which a model of a given layer is an instance of a model of the following layer, in the sense that each element of the former is an instance of an element of the latter [15]. In the current specification of the metamodeling language MOF (Meta Object Facility) [8], the OMG describes four layers as follows:

- The *information layer* (M_0) is comprised of the data that we wish to describe.
- The *model layer* (M_1) is comprised of the metadata that describes data in the information layer. Metadata is informally aggregated as models.
- The *metamodel layer* (M_2) is comprised of the descriptions (i.e., meta-metadata) that define the structure and semantics of metadata. Meta-metadata is informally aggregated as metamodels.
- The *meta-metamodel layer* (M_3) is comprised of the description of the structure and semantics of meta-metadata.

In terms of this architecture, the standard ERP development starts by creating a reference model corresponding to the *model layer* (M_1). The alternative approach proposed in this paper involves implementing a reference metamodel corresponding to the M_2 layer, named the *Enterprise Metamodel*, which is defined as an instance of MOF (as the OMG defines the UML). The *Enterprise Metamodel* encompasses an *Entity Metamodel* and a *Relationship Metamodel*, which are described in the next subsections.

3.2 Entity Metamodel

The *Entity Metamodel* (Figure 2) comprises a hierarchy of *Enterprise_Entity*, each of which is *characterized* through the assignment of a set of *Enterprise_Feature*, and the corresponding *Enterprise_Object* (for the sake of legibility, from now onwards we will omit the “Enterprise_” preceding the elements of the metamodel). The hierarchy is established by the *generalization* class, and allows feature inheritance from parent entities (general) to child entities (specific).

For each instance of *Entity* there will be multiple instances of *Object*, which will be characterized by assigning specific *Values* to the corresponding instances of *Characterization*. Therefore, features can conceptually exist independently from the entities. However, as shown in Figure 2, the feature instances only exist in connection with a specific instance of an *Object*, never autonomously. On the other hand, values have their own identity. Since, in the proposed approach, it is the metamodel that gets implemented in code, it is possible to assign data types defined at runtime to the

features. A relevant aspect, quite common in object orientation, is the possibility that one of the features in a class has a predefined value (*Specific_Characterization*), affecting all objects in that class and those in the class tree inheriting from it. From the standpoint of the proposed model, this implies that there are two types of *Characterization*, namely generic and specific.

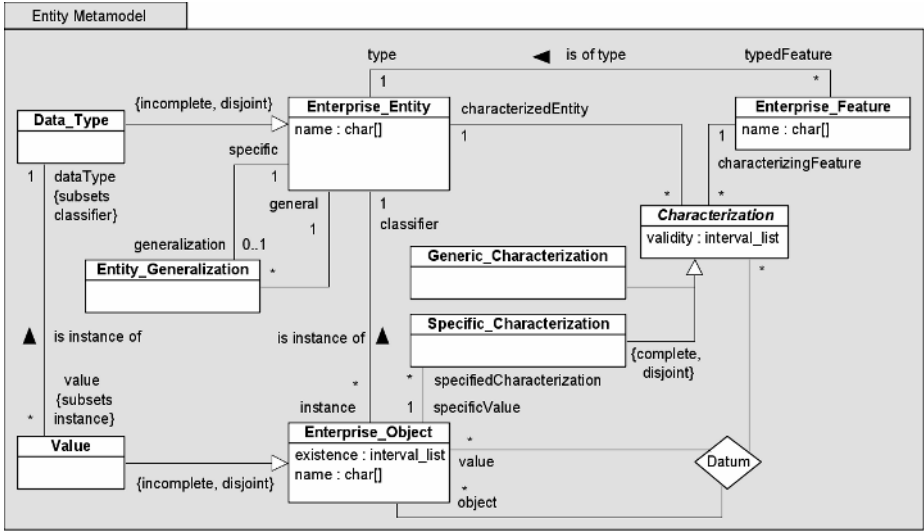


Fig. 2. Entity Metamodel package

3.3 Relationship Metamodel

The other essential elements of an enterprise model are the relationships between entities, established with commercial or management purposes. As Fig. 3 shows, the relationships package is modeled following the same approach as the *Entity Metamodel*.

The *Participation* element links the entities to the relationships, thus playing a similar role to that played by the *Characterization* element by linking features to entities. There is also a complete classification in generic and specific participations. In the *Generic_Participation* it is possible to define cardinality. The proposed metamodel conceives the enterprise relationship as a group of entities with a purpose, thus it does not require the details of the internal relationships between those entities.

The *Relationship* inheritance tree, set through the *generalization* element, allows the gradual and intuitive definition of the enterprise activity. For example, it is possible to create a “Sell” Relationship, involving the basic entities (like “Customer”, “Product”, “Sales rep”), and then to create the subtypes “Sell_Perishable” (in which the entity “Product” is replaced by its subtype “Perishable”), “Sell_Drink” (replacing “Product” by “Drink”), and so on.

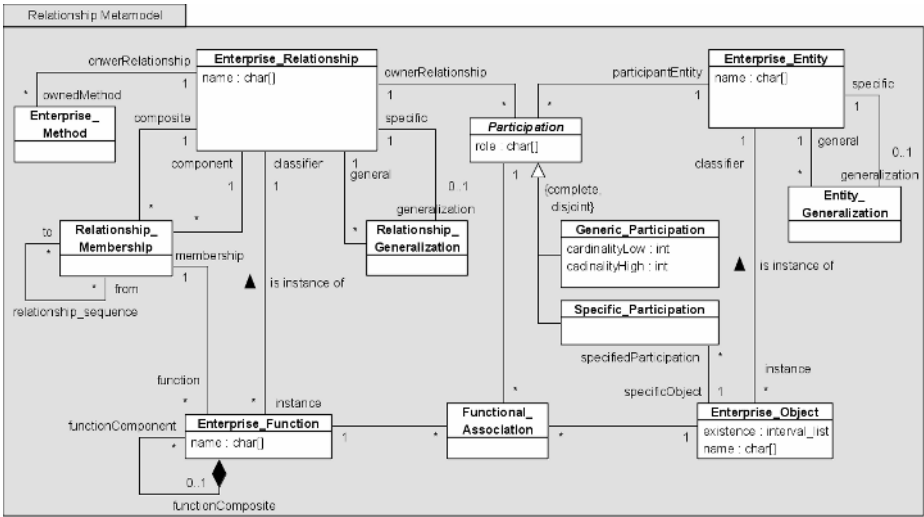


Fig. 3. Relationship Metamodel package

There is also a functional hierarchy of Relationships, with an associated sequence, so that the software can interpret the logic of the day-to-day processes. For example, the relationship “Sell” might consist of the sequence of relationships “Customer Order” -> “Invoice” -> “Ship”. The logic of the relationship resides also in the *Methods*, which result from the need to perform arithmetical operations involving the values of the features of the entities that participate in a relationship; for example, the calculation of the tax (VAT) of a sale as a fixed percentage of the sale amount. Finally, the activity of the enterprise will materialize in successive instances of *Function* (for example “Sell_Drink number #####”).

3.4 Enterprise Metamodel

The enterprise metamodel derives from the merging of the entity and the relationship packages, besides the fundamental linking of both through the central *Datum* association (Fig. 4). *Datum* becomes an essential element of the metamodel: embodies the transactional activity of the enterprise; when functions are executed, the characterization of the objects will consequently vary.

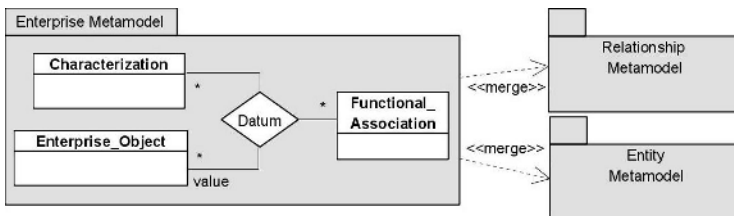


Fig. 4. Enterprise Metamodel

3.5 Metamodel Hierarchy Example

Figure 5 shows a simplified example, in which the positioning of the proposed metamodel with respect to the four-layer metamodel architecture is illustrated. As mentioned before, the metamodel is an instance of MOF. The relationships between layers are represented by the “instance of” links, while the intra-layer relationships between object instances and entity instances are represented by the “snapshot” links, which cross the logical boundary that delimits the “model zone” and the “data zone”. That resembles the equivalent diagram of the UML Infrastructure specification [15]. Please note that not all the links and labels are included and that the metamodel is incomplete; otherwise the picture would become unclear.

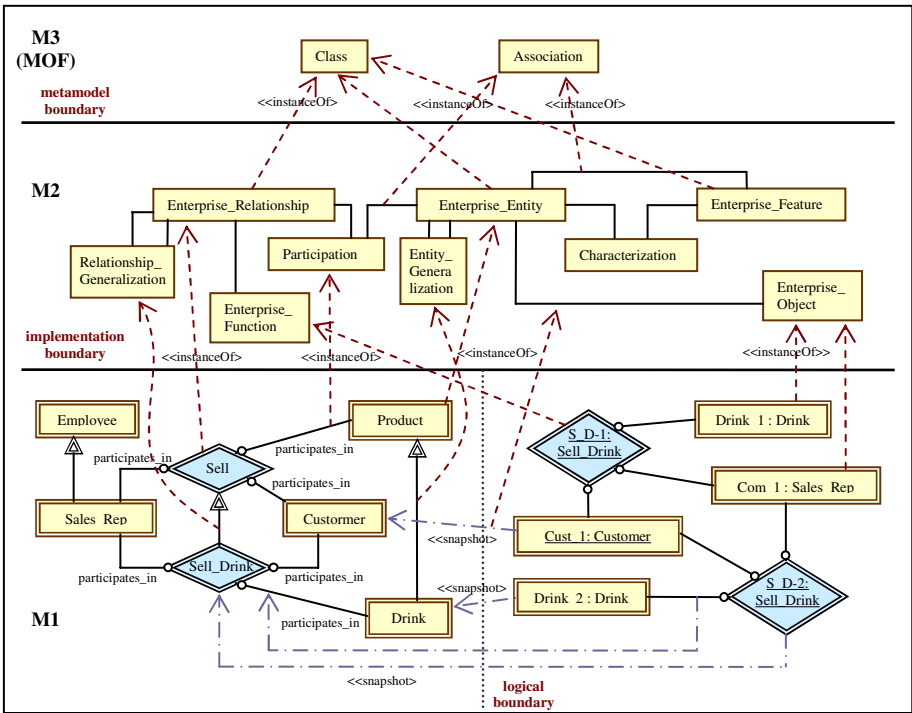


Fig. 5. Positioning of the metamodel example

In order to clarify the semantics of the metamodel, a simple model, which is an instance of the relationship package, is depicted in layer M1. Some conventions are adopted for the representation of enterprise models, taking UML as a reference. The instances of Entity are represented as double-lined boxes; the Generalization association is represented by a line ending in a double-lined triangle; the Relationships are represented as double-lined diamonds; the Participation instances are circle-ended lines linking entities to relationships, and introducing the expression “participates in” above the line; finally, all the object-level instances are represented, like in UML, by underlining the name and referring to the respective entity.

4 Alternative Development Cycle

Based on all the concepts developed throughout the paper, this section summarizes the alternative enterprise software development cycle —resulting from the implementation of the metamodel presented in the third section— represented by the sequence of stages depicted in Figure 6:

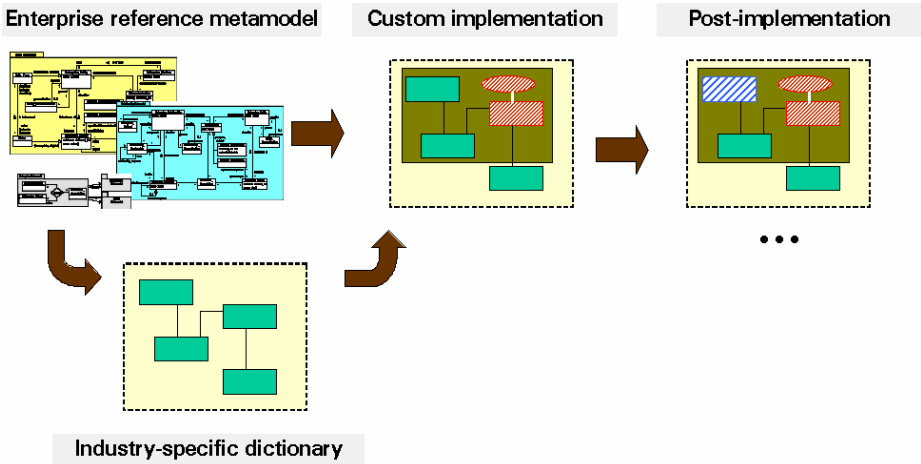


Fig. 6. The alternative ERP life cycle based on enterprise metamodeling

- **Enterprise reference metamodel.** The starting point for the alternative cycle is the implementation in a database of the proposed metamodel of enterprise entities and relationships. This approach implies a higher generality as well as a flexible capability to adapt to specific requirements.
- **Industry-specific dictionary.** Starting a company definition from scratch would be very effort-consuming (even though feasible). However, there is a basic hierarchy of entities and relationships essentially common to all enterprises in any given sector. Thus, the role played in the ERPs by the module-based and industry-specific solutions is performed, in the proposed approach, by a set of database pre-loads, containing a pre-definition of entities, features and relationships.
- **Custom implementation.** A remarkable trait of the alternative approach is that the result of the adaptation of the model to the company is already customized (in the figure, it is depicted as identical to the customized model of the traditional approach). Therefore, the ad-hoc customization costs are avoided. Additionally, the user company might find it easier to identify itself with the customized entity hierarchy, which is closer to its reality, thus facilitating its involvement and consequently improving the requirement identification.
- **Post-implementation.** Finally, the alternative approach provides a special flexibility in the post-implementation stage, by facilitating the adaptation to changing requirements since there is no pre-coded enterprise model, but a pre-coded enterprise metamodel.

5 Experimentation

The practical application of the proposed alternative development cycle has established its potential to achieve flexibility in packaged enterprise software for small and medium enterprises (SMEs). A first version of an enterprise management software package developed following this approach is currently being utilized by several SMEs; it has also been recommended by the Spanish reprographic association to its member companies due basically to its flexibility to implement ad-hoc requirements.

The current version of the software package is oriented to small companies searching for low cost software. It is database oriented software, with client/server architecture, with most of the code residing in the database (programming SQL). The client applications are windows-based; essentially are GUIs (Graphical User Interfaces), developed in Delphi, which access to the server database through TCP/IP protocol. It runs over either Linux or MS Windows operating system, uses the open source database FireBird 1.5.2 and has low hardware requirements (Server: Intel Pentium/AMD 3GHz and 1Gb RAM; Client: Intel Pentium/AMD 1GHz and 256Mb RAM). It encompasses two main modules:

- **ArmorArqt.** This is the module used to model the enterprise, i.e., to create and characterize the entities and to create and define the participations of the relationships. ArmorARQT is used to populate the database that implements the metamodel. An inherent characteristic of a metamodel is the intrinsic definition of a language (Domain-Specific Language (DSL)) [13], [16].
- **ArmorPrise.** This is the enterprise management software. It follows a generic logic that starts by identifying the main relationships (those which are the first level components of the master relationship which is the enterprise itself) and grouping them in a menu option. The software then looks for the component relationships of each one. The user selects the relationship to be executed. The corresponding function is then created, and the software asks for the required objects as defined in the participations of the relationship. The methods associated with the relationship can be called to execute the defined operations.

6 Conclusions

This paper presents an enterprise reference metamodel whose implementation leads to an alternative viable ERP life cycle. The proposed approach offers greater flexibility to accommodate company requirements than the current ERP development approach based on reference models.

There is a noticeable improvement in the implementation stage, since the user of the package can be more involved in the modeling of its company, which is carried out in a much more intuitive fashion, thus improving the identification of requirements. The improvement is also substantial in the post-implementation stage, given the possibilities provided by the proposed flexible approach to modify the initial company model and to accommodate the series of changes in the requirements that normally arise after the go live date. Furthermore, this flexibility encourages the company to execute more frequent low-cost adaptations. To conclude, the experimentation with a real implantation of a software package developed according to the

approach proposed in this paper, suggests that the customized result of the implementation and the subsequent flexibility are particularly advantageous, since ad-hoc development costs are drastically reduced as compared with the traditional approach.

Acknowledgements

This paper has been prepared in the context of *ARMORsystem*, a collaborative research project between the ADALID Co. and the Carlos III University of Madrid, with financial support from FEDER, CDTI and the Instituto de Fomento de Murcia.

References

1. Davenport, T.H.: Putting the Enterprise into the Enterprise System. *Harvard Business Review* 76 (4) (1998) 121–131
2. Rosemann, M., van der Aalst, W.M.P.: A configurable reference modeling language. *Information Systems*, available on-line, to appear in (2006).
3. Scheer, A.-W.: *Business Process Engineering: Reference Models for Industrial Enterprises*. 2nd edn. Springer-Verlag, Berlin [etc.] (1994)
4. Brehm, L., Heinzl, A., Markus, M.L.: Tailoring ERP Systems: A Spectrum of Choices and their Implications. In: *Proceedings of the 34th Hawaii International Conference on System Sciences*, Vol. 8. IEEE (2001) pap. 17
5. Luo, W., Strong, D.M.: A Framework for Evaluating ERP Implementation Choices. *IEEE Transactions on Engineering Management* 51 (3) (2004) 322–3333.
6. Soffer, P., Golany, B. Dori, D.: Aligning an ERP System with Enterprise Requirements: An Object-Process Based Approach. *Computers in Industry* 56 (6) (2005) 639–662
7. Scheer, A.-W., Habermann, F.: Making ERP a Success. *Communications of the ACM* 43 (4) (2000) 57–61
8. OMG: Meta Object Facility (MOF) Specification, Version 1.4. (2002) 2-1–2-5. Downloadable at <http://www.omg.org>
9. SAP: Solution Composer: What is it? SAP AG (2005). Downloadable at <http://www.sap.com/solutions/businessmaps/composer/index.epx> (accessed oct. 2005)
10. SAP: Designed for Your Industry, Scaled to Your Business, Ready for Your Future. SAP AG (2004). Downloadable at <http://www.sap.com/industries/index.epx> (accessed dec. 2005)
11. Oracle: Oracle Industries Solutions. <http://www.oracle.com/industries> (accessed dec. 2005)
12. Soffer, P., Golany, B. Dori, D.: ERP modeling: a comprehensive approach. *Information Systems* 28 (6) (2003) 673–690
13. Petit, M. (ed.), Domeingts, G. (approval): Report on the State of the Art in Enterprise Modelling. UEML Deliverable D1.1 (2002). <http://www.ueml.org>
14. Tannenbaum, A.: *Metadata Solutions: Using Metamodels, Repositories, XML and Enterprise Portals to Generate Information on Demand*. Addison Wesley, Boston [etc.] (2002)
15. OMG: Unified Modeling Language (UML) Specification: Infrastructure, version 2.0, ptc/04-10-14. (2004) Downloadable at <http://www.omg.org/>
16. Kovse, J., Weber, C., Härder, T.: Metaprogramming for Relational Databases. In: Atzeni, P., Chu, W.W., Lu, H., Zhou, S., Ling, T.W. (eds.): *Conceptual Modeling - ER 2004. Lecture Notes in Computer Science*, Vol. 3288, Springer-Verlag, Berlin Heidelberg New York (2004) 654–667

Static Analysis Based Software Architecture Recovery

Jiang Guo, Yuehong Liao, and Raj Pamula

Department of Computer Science, California State University Los Angeles,
Los Angeles, California, USA
{jguo, yliao2, rpamula}@calstatela.edu

Abstract. Recover the software architectures is a key step in the reengineering legacy (procedural) programs into an object-oriented platform. Identifying, extracting and reengineering software architectures that implement abstractions within existing systems is a promising cost-effective way to create reusable assets and reengineer legacy systems. We introduce a new approach to recover software architectures in legacy systems. The approach described in this paper concentrate especially on how to find software architectures and on how to establish the relationships of the identified software components. This paper summarizes our experiences with using computer-supported methods to facilitate the reuse of the software architectures of the legacy systems by recovering the behavior of the systems using systematic methods, and illustrate their use in the context of the Janus System.

1 Introduction

Software industry has a lot of legacy programs needing reengineering and modernization. One modernization approach is to convert old (procedural) programs into an object-oriented platform to make them easier to understand, maintain and reuse. One important purpose of the software reengineering is to develop large software systems built on several legacy systems to make use of the partial or full functionalities of these legacy systems [1]. A critical issue for reengineering the legacy systems is the identifying, extracting, modeling and analysis of software architectures. Software architecture is the nature of interactions among the software components. The software architecture concerns the design of the gross structure of a software system, including its overall behavior and its decomposition in simpler computational elements. An architectural description singles out the components from which a system is built, describing their functional behavior and providing a complete description of their interactions. The use of explicit descriptions of the architecture of software systems enhances system comprehension and promotes software reuse. Besides, software architecture helps verify the structural properties of the system to be developed.

Modernizing the architecture of old software helps to gain control over maintenance cost and to improve system performance, while supporting the move to a distributed or more efficient environment [2]. Studies by Gall [3] also found that a

re-architecting of old procedural software to object-oriented architecture results in object-oriented software that helps reduce future maintenance cost, since modern maintenance technology can then be applied. Thus, when converting procedural programs into object-oriented ones, we must identify the software architecture and objects that are reusable from procedural code. In the software architecture identification phase, we need reengineering means. The methods described in this paper concentrate especially on how to find software architecture and objects (classes) and on how to establish the relationships of the identified classes.

The essence of software reengineering is to improve or transform existing software so that it can be understood, controlled, and used anew. The need for software reengineering has increased greatly, as heritage software systems have become obsolescent in terms of the platforms on which they run, and their suitability and stability to support evolution to support changing needs [4]. Software reengineering is important for recovering and reusing existing software assets, putting high software maintenance costs under control, and establishing a base for future software evolution [5]. Identifying, extracting and reengineering software architecture and components that implement abstractions within existing systems is a promising cost-effective way to create reusable assets and reengineer legacy systems [6].

The research was motivated by the need for better techniques for the extraction and utilization of desirable functionality of an existing system for reengineering, reuse, and maintenance. We present a new program slicing process for identifying and extracting software architectures implementing functional abstractions. Once the slicing criterion has been identified the slice is isolated using algorithms based on dependence graphs. Both symbolic execution and program slicing are performed by exploiting the Data Flow Graph (DFG) and Control Flow Graph (CFG), a fine-grained dependence based program representation that can be used for software architecture recovery tasks. The work described in this paper is aiming to explore reverse engineering and reengineering techniques for reusing software architecture from existing systems.

2 Related Work

Developing complex software systems requires a description of the structure or structures, which comprise software components, the externally visible properties of those components, and the relationships among them [7]. Such a description, called software architecture, also is basic for further engineering activities concerning reuse, maintenance, and evolution of existing software components and systems.

Architecture recovery refers to all techniques and processes used to abstract a higher-level representation (i.e., software architecture) from available information such as existing artifacts (e.g., source code, profiling information, design documentation) and expert knowledge (e.g., software architects, maintainers) [8]. Basically this means the extraction of those building blocks that constitute architectural properties and finally the software architecture. From point of this view we think of architectural styles and patterns that are inherent in almost any design and thus are primary objectives for architecture recovery [9].

Architecture recovery has received considerable attention recently and various frameworks, techniques and tools have been developed [10]. Basically, existing knowledge, obtained from experts and design documents, and various tools are mandatory to solve the problem. Hence, a common idea is to integrate several tools in architecture workbenches such as Dali [11]. In this a variety of lexical-based, parser-based and profiling-based tools are used to examine a system and extract static and dynamic views to be stored in a repository. Analyses of these views are supported by visualization and specific analysis tools. They enable an interaction with experts to control the recovery process until the software architecture is reconstructed. Concerning architecture reconstruction much work has been on techniques that combine bottom-up and topdown approaches. Fiutem et al. describe such an approach [12]. They use reverse engineering tools to extract source models (e.g., Abstract Syntax Tree) and top-down they apply queries to extract expected patterns. They use a hierarchical architectural model that drives the application of a set of recognizers. Each recognizer works on the Abstract Syntax Tree (AST) and is related to a specific level of the architectural model. They produce different abstract views of the source code that describe some architectural aspects of the system and are represented by hierarchical architectural graphs. Harris et al. outline a framework that integrates reverse engineering technology and architectural style representations [13]. In bottom-up recovery the birds eye view is used to display the file structure and file components of the system, and to reorganize information into more meaningful clusters. Top-down style definitions place an expectation on what will be found in the software system. These expectations are specified by recognition queries that are then applied to an extracted AST. Each recognized style provides a view of the system and the collection of these views partially recovers the overall design of the software system. Guo et al. outline an iterative and semi-automatic architecture recovery method called ARM [14]. ARM supports patterns at various abstraction levels and uses lower-level patterns to build higher-level patterns and also composite patterns. In this way the approach aims particularly at systems that have been developed using design patterns whose implementations have not eroded over time. Another approach which uses source models and queries as basic inputs for architecture recovery is introduced by Sartipi et al. and called Alborz [15]. The problem is viewed as approximate graph matching problem whereas the extracted source models and defined queries are represented as attributed relational graphs.

3 Slicing-Based Static Analysis

We start the reengineering process by extracting the requirements and detailed design information from the source code and existing documents. It is the process of analyzing a subject system to identify the system's components and their interrelationships and create software architecture representations of the system in another form or at a higher level of abstraction and expressed using data flow and control flow diagrams.

The goal of source information extraction is to enable the recovery of the software architecture. We use statically extracting and dynamically extracting. A significant quantity of information may be extracted from the static artifacts of software systems, such as source code, makefiles, and design models, using techniques that include parsing and lexical analysis.

We initiated the software architecture recovery process with a preprocessing step that restructures code. We built on the theory that unstructured code can be written using only D-structures [16] and relied on existing algorithms for that purpose [17]. Our research within this phase involves the use of program slicing techniques for isolating code fragments implementing functional abstractions.

Program slicing is a static analysis technique that extracts all statements relevant to the computation of a given variable. This is accomplished by using data-flow analysis to analyze the program source code without the need to actually execute the program. Program slicing, an application of data-flow analysis, can be used to transform a large program into a smaller one containing only those statements relevant to the computation of a given variable. For example someone is interested in how the value for the out parameter *r2* is computed in the following program:

```

procedure myproc(a, b, c: integer; r1, r2: out integer)
  v, w : integer := 0;
  begin
    for i in 1..a loop
      w := w + b;
    end loop;
    for i in 1..c loop
      v := i*b + c + v;
    end loop;
    r1 := w;
    r2 := v*2;
  end myproc;

```

All we need to know is the following program slice which is obviously easier to understand:

```

procedure myproc( ... b, c: integer; ... r2: out integer)
  v, ... : integer := 0;
  begin
    ...

    for i in 1..c loop
      v := i*b + c + v;
    end loop;

    ...
    r2 := v*2;
  end myproc;

```

A Program Slice is defined as follows: Given a syntactically correct source program P , in some programming language, and a slicing criterion $C = \langle L, V \rangle$. Where L is a location in the program and V is a variable in the program. S is a slice of program P for criterion C if (1) S is derived from P by deleting statements from P , (2) S is syntactically correct, and (3) for all executions of P and S , in any given execution of P and of S with the same inputs, the value of V in the execution of slice S just before control reaches location L is the same as the value of V in program P just before control reaches location L .

The function of the slicing criterion is to specify the program variable that is of interest along with a location in the program where the value of the variable is desired.

Our program slicing tool constructs program slices from the control structure of the program and the pattern of assignment and reference to variables by backward chaining from the slicing criterion to the beginning of the program.

The following definitions are helpful in understanding how program slices are constructed.

$Defs(n)$: The set of variables defined (assigned to) at statement n .

$Refs(n)$: The set of variables referenced at statement n .

$Reqs(n)$: A set of statements that is included in a slice along with statement n . The set is used to specify control statements (e.g., if or while) enclosing statement n or other characters that are syntactically part of statement n but are not contiguous with the main group of characters comprising the statement.

An algorithm for constructing program slices must locate all statements relevant to a given slicing criterion. The essence of a slicing algorithm is the following: starting with the statement specified in the slicing criterion, include each predecessor statement that assigns a value to any variable in the slicing criterion, generate a new slicing criterion for the predecessor by deleting the assigned variables from the original slicing criterion, and add any variables referenced by the predecessor.

For expression statement n , a predecessor of statement m , the $Defs(n)$ set and the slicing criterion determines if an expression statement is included in a slice.

$$S_{\langle m, v \rangle} = \begin{cases} \{n\} \cup S_{\langle n, v \rangle} & \forall x \in Refs(n) \text{ if } v \in Defs(n) \\ S_{\langle n, v \rangle} & \text{otherwise} \end{cases} \quad (1)$$

A compound control statement is a statement that has a condition directly controlling the execution of another statement (possibly also a compound statement). Control statements such as *if*, *switch*, *while*, *for* and *do...while* should be included in a program slice whenever any statement governed by the control statement is included in a slice. When control statement n is added to a program slice, the slice on the criterion $\langle n, Refs(n) \rangle$ is added to the original slice. For each statement, n , associate a set, $Reqs(n)$, of statements that are required to be included in any slice containing statement n . The slicing rule for $v \in Defs(n)$ becomes:

$$S_{\langle m, v \rangle} = \{n\} \cup \left(\bigcup_{x \in Refs(n)} S_{\langle n, x \rangle} \right) \cup \left(\bigcup_{y \in Refs(k)} \bigcup_{k \in Reqs(n)} S_{\langle k, y \rangle} \right) \quad (2)$$

The result of the revised rule is to include the set of required statements for statement n , $Reqs(n)$, whenever statement n is included in a slice. Where, x, y are referenced variables at statement n and k . k is a statement included in a slice along with statement n . From unions and intersections of slices, a slice-based model of program structure that has applications to program understanding tasks can be built.

Program slicing has been used both as structural and specification driven method. As structural method, program slicing has been used to identify external user functionalities in large programs. The isolation of an internal domain dependent function can be driven by its formal specification. The specification can be used together with symbolic execution techniques to identify a suitable slicing criterion. Code segmentation is needed in order to reduce the granularity and thus the complexity of the remaining processes. We have defined a segmentation scheme that separates the code into modular units while also removing syntactic sugar features of the code. We have also defined heuristics to attach in-code documentation to the appropriate segment. For a program P the result is a set of segments, such that $SG = \{sg_1, sg_2, \dots, sg_n\}$ and $P_f = \bigcup sg_i$, where $1 < i < n$ and P_f represents code that is identical in functionality to P and sg_i is a segment.

Following the segmentation, we defined dependency algorithms that analyze each sg_i . Specific slicing algorithms that are modified forms of the slicing algorithms [18] are employed at the statement, construct, and block levels. These algorithms provide information on all variables, which is the start point of the software architecture recovery.

4 Software Architecture Recovery

Right now, most software architecture recovery and reengineering approaches have in common that they all take into account patterns to reconstruct the architecture of a software system. We regard software components (functions, modules, etc.) and their relations as the key elements of software systems residing in all levels of abstraction. Thereby we start software element and relation recognition from the lowest level (i.e., source level) and use hot-spots to stepwise abstract higher-level patterns. Hot-spots indicate patterns and are represented by meaningful source code structures (e.g., variables, functions, data structures, program structures). To detect such hot-spots in source code we apply extended string pattern matching which facilitates fast and effective queries.

The important step in the transformation of the code and software architecture is to reconstruct the existing software architecture. This is done by extracting architecturally important features from its code and employing architectural reconstruction rules to aggregate the extracted (low-level) information into an

architectural representation [19]. This process is to extract the systems architecture, we performed a static extraction of the source code to identify function-call and built-from relations. The former identifies function calls within the system. The latter identifies how the executables in the system are built from the source files. In addition, we performed dynamic extraction to determine the run-time connectivity of the system as a whole.

Based on the static analysis of the program slicing and the results collected from running the instrumented system, we can reconstruct the run-time connectivity of the architecture. We can construct a control-dependence graph (CDG). The CDG, in its most basic form, is a Directed Acyclic Graph (DAG) that has program predicates as its root and internal nodes, and nonpredicates as its leaves. Nodes in CDG may be basic blocks and statements. A leaf is executed during the execution of the program if the predicates on the path leading to it in the control-dependence graph are satisfied. More specifically, let $G = \langle N, E \rangle$ be a flow graph for a procedure. A node m postdominates node n , written $m \text{ } pdom \text{ } n$, if and only if every path from n to $Exit$ passes through m . Then node n is control-dependent on node m if and only if (1) There exists a control-flow path from m to n such that every node in the path other than m is postdominated by n and (2) n does not postdominate m .

The results of the restructuring, segmentation, and dependency steps are segment design representations and a global design representation. These representations include traditional methods, such as call graphs, structure charts, and hierarchical diagrams and other less conventional representations such as variable usage and state change descriptions (such as state-chart). These representations serve as input to perform object identification and to create formal specifications of object behavior. Results of our work include methods that recover the design information at varying levels of granularity, expressible in numerous forms from both data and functional viewpoints. The data and control dependency representations are the basis for our software architecture recovery research.

Once the architecture has been extracted, we can view the components and connectors of the system as possessing architecturally significant properties. We listed the possible features of architectural elements (both components and connectors), and are divided into information that can be derived from a temporal perspective and information that can be derived from a static perspective of an architectural element.

With the extracted information, we can reconstruct the software architecture with the relation partition algebra. A set is a collection of objects, called elements or members. If x is an element of S , given any object x and set S , we write $x \in S$. The notion of set and the relation is-element-of are the primitive concepts of set theory. A finite set can be specified explicitly by enumerating its elements. We use set to represent the software architecture components of a system. For example,

$$\textit{Subsystems} = \{OS, Drivers, DB, Application\}$$

$$\textit{Functions} = \{main, a, b, c, d\}$$

$$\textit{InitFunctions} = \{f \mid f \in \textit{Functions} \wedge f \text{ is called at initialization time } \}$$

and we also have

$InitFunctions \subseteq Functions$.

Relationships between software components play an important role in software architecture and design. Binary relations can express such relationships. For example, function-calls within a system can be seen as the binary relation named *calls*, such as $calls(main, a)$ is an abstraction of the main program *main* calls a function *a*.

Then, we can represent a software architecture in a directed graph. A directed graph consists of a set of elements, called vertices, and a set of ordered pairs of these elements, called arcs.

5 Case Study and Verification

We have explored software architecture recovery in the context of a case study that addresses the reengineering of the Janus System. Janus is a software-based war game that simulates ground battles between up to six adversaries. Janus is interactive in that command and control functions are entered by military analysts who decide what to do in crucial situations during simulated combat. The current version of Janus operates on a Hewlett Packard workstation and consists of a large number of FORTRAN modules, organized as a flat structure and interconnected with one another via FORTRAN COMMON blocks. This software structure makes modification of Janus very costly and error-prone. There is a need to modernize the Janus software into a maintainable and evolvable system and to take advantage of modern personal computers to make Janus more accessible to the Army.

The first step in our process, system and requirements understanding, took the form of a series of brief meetings with the client, which also included a short demonstration of the current software system. Our goal was to gather as much information as we could about the currently existing system to aid in gaining a clearer understanding of its present functionality.

Next, we proceeded to develop object models of the Janus System using the aforementioned materials and products, to create the modules and associations amongst them. This was probably the most difficult and most important step. It required a great deal of analysis and focus to transform the currently scattered sets of data and functions into small, coherent and realizable objects, each with its own attributes and operations.

We used our approach to reuse the information extracted from the old system. The most important type of reuse was reuse of implicit domain models. We reused the domain analysis and knowledge since the domain was stable across the reengineering transformations. This greatly reduced the time and effort that needed be spent on domain related work, such as the analysis of the domain dependent functions. Second was reuse of software architecture. This kind of reuse included the user functionalities, functional abstraction, task flow, and user interface specifications. Third was the reuse of data models. The reuse of

data models was very helpful to re-organize the data information although we needed to transform the old data structures into new data structures. Fourth was the reuse of algorithms. The code could not be reused directly because it had to be transformed into another language (Ada). However, the main algorithms were the same - we did not need to redesign the algorithms, we just rewrote them in new languages.

Based on the feedback of our object models from the domain experts, the reengineering team revised the object models for the Janus core elements and developed a 3-tier object-oriented architecture for the Janus System. The new architecture of Janus uses an explicit priority queue of event objects to schedule the simulation events.

6 Conclusion

Discover the objects and software architecture from the legacy systems is the essence of software reengineering. The focus of the reengineering effort was to abstractly capture the systems functionality and then produce system models that would most accurately represent that functionality, while factoring out independent concerns and aspects that were likely to change. Large systems are divided into subsystems. These subsystems, also known as components and objects, and the dependencies that exist among the components and objects form the different layers within a software architecture. In this paper, we propose an approach to extract the software architecture based on the programming slicing and parameters analysis and dependencies between the objects based on the relation partition algebra.

Our future work will focus on automating the approach of extracting reusable software components from legacy systems and deriving the software architecture based on the dependencies of the components.

References

1. Favre, J., Sanlaville, R., Continuous Discovery of Software Architecture in a Large Evolving Company, Workshop on Software Architecture Reconstruction, the Working Conference on Reverse Engineering (2002)
2. Hall, P., Architecture-driven Component Reuse, Information and Software Technology, Vol. 41, Issue 14 (1999)
3. Gall, H., R. Klsch, R. Mittermeir, Application Patterns in Reengineering: Identifying and Using Reusable Concepts, Proceedings of the 5th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Granada, Spain (1996)
4. Ali, F., Du, W., Toward reuse of object-oriented software design models. Information and Software Technology, Vol. 46 Issue 8 (2004)
5. Hakala, K., Hautamki, J., Koskimies, K., Annotating Reusable Software Architectures with Specialization Patterns Proceedings of the Working IEEE/IFIP Conference on Software Architecture (2001)

6. Guo, J., Towards Semi-Automatically Reusing Objects from Legacy Systems, *International Journal of Computers and Their Applications*, Vol. 11, No. 3 (2004)
7. Goseva-Popstojanova, K., Trivedi, K., Architecture-Based Approaches to Software Reliability Prediction, *Computers and Mathematics with Applications*, Vol. 46 (2003)
8. Svetinovic, D., Godfrey, M., A Lightweight Architecture Recovery Process, *Software Architecture Recovery and Modelling*, Stuttgart, Germany (2001)
9. Ali-Babar, M., Zhu, L., Jeffery, R., A Framework for Classifying and Comparing Software Architecture Evaluation Methods," *Australian Software Engineering Conference*, Melbourne (2004)
10. Pinzger, M., Gall, H., Pattern-Supported Architecture Recovery, *Proceedings of 10th International Workshop on Program Comprehension*, IEEE Computer Society Press, Paris, France (2002)
11. Kazman, R., Carriere, S., View Extraction and View Fusion in Architectural Understanding, *Proceedings of the 5th International Conference on Software Reuse*, Victoria, BC, Canada (1998)
12. Fiutem, R., Tonella, A., Antonioli, G., Merlo, E., A Clich-based Environment to Support Architectural Reverse Engineering, *Proceedings of the International Conference on Software Maintenance*, Monterey, California (1996)
13. Harris, D., Reubenstein, H., Yeh, A., Reverse Engineering to the Architectural Level, *Proceedings of the 17th International Conference on Software Engineering*, Seattle, Washington (1995)
14. Guo, G., Atlee, J., Kazman, R., A Software Architecture Reconstruction method, *Proceedings of the 1st Working IFIP Conference on Software Architecture*, San Antonio, Texas, (1999)
15. Sartipi, K., Kontogiannis, K., Mavaddat, F., A Pattern Matching Framework for Software Architecture Recovery and Restructuring, *Proceedings of the 8th International Workshop on Program Comprehension*, Limerick, Ireland (2000)
16. Dijkstra, E., *A Discipline of Programming*, Prentice Hall (1976)
17. Boehm, C., Jacopini, G., Flow Diagrams, Turing Machines, and Languages with only Two Formation Rules, *Communications of the ACM*, Vol. 9 No. 5 (1966)
18. Atkinson, D., Griswold, W., Implementation Techniques for Efficient Data-flow Analysis of Large Programs, *Proceedings of the International Conference on Software Maintenance* (2001)
19. Carriere, S., Woods, S., Kazman, R., Software Architectural Transformation, *Proceedings of 6th Working Conference on Reverse Engineering*, Atlanta, Georgia (1999)

A First Approach to a Data Quality Model for Web Portals

Angelica Caro¹, Coral Calero², Ismael Caballero², and Mario Piattini²

¹ Universidad del Bio Bio, Departamento de Auditoria e Informática,
La Castilla s/n, Chillán, Chile
mcaro@ubiobio.cl

² ALARCOS Research Group,
Information Systems and Technologies Department,
UCLM-Soluziona Research and Development Institute,
University of Castilla-La Mancha, Paseo de la Universidad, 4 – 13071 Ciudad Real, Spain
{Coral.Calero, Ismael.Caballero, Mario.Piattini}@uclm.es

Abstract. The technological advances and the use of the internet have favoured the appearance of a great diversity of web applications, among them Web Portals. Through them, organizations develop their businesses in a really competitive environment. A decisive factor for this competitiveness is the assurance of data quality. In the last years, several research works on Web Data Quality have been developed. However, there is a lack of specific proposals for web portals data quality. Our aim is to develop a data quality model for web portals focused on three aspects: data quality expectations of data consumer, the software functionality of web portals and the web data quality attributes recompiled from a literature review. In this paper, we will present the first version of our model.

1 Introduction

In the last years, a growing interest in the subject of Data Quality (DQ) or Information Quality (IQ) has been generated because of the increase of interconnectivity of data producers and data consumers mainly due to the development of the internet and web technologies. The DQ/IQ is often defined as “fitness for use”, i.e., the ability of a data collection to meet user requirements [1, 2]. Data Quality is a multi-dimensional concept [2], and in the DQ/IQ literature several frameworks providing categories and dimensions as a way of facing DQ/IQ problems can be found.

Research on DQ/IQ started in the context of information systems [1, 3] and it has been extended to contexts such as cooperative systems [4-6], data warehouses [7, 8] or electronic commerce [9, 10], among others.

Due to the characteristics of web applications and their differences from the traditional information systems, the community of researchers has recently started to deal with the subject of DQ/IQ on the web [11]. However, there are not works on DQ/IQ specifically developed for web portals. As the literature shows that DQ/IQ is very dependent on the context, we have centred our work on the definition of a Data Quality Model for web portals. To do so, we have used some works developed for differ-

ent contexts on the web but that can be partially applied or adapted to our particular context. For example, we have used the work of Yang et al., (2004) where a quality framework for web portals is proposed including data quality as a part of it.

As the concept of “fitness for use” is widely adopted in the literature (emphasizing the importance of taking into consideration the consumer viewpoint of quality), we have also considered, for the definition of our model, the data consumer viewpoint. First, we have combined the data quality expectations of data consumers with the software functionality of web portals. From the resultant matrix (data consumer expectations x functionalities), we have determined which web data quality attributes, recompiled in a literature review, can be applied.

The structure of this paper is as follows. In section 2, the components of our model are presented. In section 3, we will deeply describe the first version of our DQ/IQ Web Portal Model. Finally, in section 4 we will conclude with our general remarks and future work.

2 Model Components

Web Portals are emerging Internet-based applications that enable access to different sources (providers) through a single interface [12]. The primary objective of a portal software solution is to create a working environment where users can easily navigate in order to find the information they specifically need to perform their operational or strategic functions quickly as well as to make decisions [13], being responsibility of web portals’ owners the achievement and maintenance of a high information quality state [14].

In this section, we will present the three basic aspects considered to define our DQ/IQ model for web portals: the DQ/IQ attributes defined in the web context, the data consumer expectations about data quality, and web portals functionalities.

2.1 Data Consumer Expectations

When data management is conceptualized as a production process [1], we can identify three important roles in this process: (1) *data producers* (who generate data), (2) *data custodians* (who provide and manage resources for processing and storing data), and (3) *data consumers* (who access and use data for their tasks).

As in the context of web-based information systems, roles (1) and (2) can be developed by the same entity [11], for web portals context we identify two roles in the data management process: (1) *data producers-custodians*, and (2) *data consumers*.

So far, except for few works in DQ/IQ area, like [1, 2, 15, 16], most of the works on the subject have looked at quality from the data producer-custodian perspective. This perspective of quality differs from this in two important ways [15]:

- Data consumer has no control over the quality of available data.
- The aim of consumers is to find data that match their personal needs, rather than provide data that meet the needs of others.

Our proposal of a DQ/IQ model for web portals considers the data quality expectations of data consumer because, at the end, it is the consumer who will judge whether a data is fitted for use or not [16].

We will use the quality expectations of the data consumer on the Internet, proposed in [17]. These expectations are organized into six categories: Privacy, Content, Quality of values, Presentation, Improvement, and Commitment.

2.2 Web Portal Functionalities

A web portal is a system of data manufacturing where we can distinguish the two roles established in the previous subsection. Web portals present basic software functionalities to data consumer deploying their tasks and under our perspective, the consumer judges data by using the application functionalities. So, we used the web portal software functions that Collins proposes in [13] considering them as basics in our model. These functions are as follows: Data Points and Integration, Taxonomy, Search Capabilities, Help Features, Content Management, Process and Action, Collaboration and Communication, Personalization, Presentation, Administration, and Security. Behind these functions, the web portal encapsulates the producer-custodian role.

2.3 Web Data Quality Review

By using a DQ/IQ framework, organizations are able to define a model for data, to identify relevant quality attributes, to analyze attributes within both current and future contexts, to provide a guide to improve DQ/IQ and to solve data quality problems [18]. In the literature, we have found some proposals oriented to DQ/IQ on the web.

Among them, we can highlight those showed in table 1. Related to such proposals, we can conclude that there is no agreement concerning either the set of attributes or, in several cases, their meaning. This situation, probably, is a consequence of the different domains and author's focus of the studied works.

However, from this revision we captured several data quality attributes. The most considered are (we present between brackets different terms used for the same concept): Accuracy (Accurate), in 60% of the works; Completeness, in 50% of the works and Timeliness (Timely), in 40% of the works; Concise (Concise representation), Consistent (Consistent representation), Currency (Current), Interpretability, Relevance, Secure (Security), in 30% of the studies. Accessibility (Accessible), Amount of data (Appropriate amount of information), Availability, Credibility, Objectivity, Reputation, Source Reliability, Traceability (Traceable), Value added are stated in 20% of the works. Finally, Applicable, Clear, Comprehensive, Confidentiality, Content, Convenient, Correct, Customer Support, Degree of Duplicates, Degree of Granularity, Documentation, Understand ability (Ease of understanding), Expiration, Flexibility, Freshness, Importance, Information value, Maintainable, Novelty, Ontology, Pre-decision availability, Price, Reliability, Response time, Layout and design, Uniqueness, Validity, and Verifiability are only studied in 10 % of the works.

Summarizing the above-mentioned attributes, by means of similarity in their names and definitions, we have obtained a set of 28 attributes. Based on these DQ/IQ attributes we will try to identify which ones are applicable to the web portals context by classifying them into the matrix construed by the previous aspects (data consumer expectations x functionalities).

Table 1. Summary of web DQ/IQ framework in the literature

Author	Domain	Framework structure
[19]	Personal web sites	4 categories and 7 constructors
[20]	Data integration	3 classes and 22 of quality criterion
[10]	e-commerce	7 stages to modelling DQ problems
[9]	e-commerce	4 categories associated with 3 categories of data user requirements.
[21]	Web information systems (data evolution)	4 categories, 7 activities of DQ design and architecture to DQ management.
[6]	e-service cooperative	8 dimensions
[22]	Decision making	8 dimensions and 12 aspects related to (providers/consumers)
[23]	Web sites	4 dimensions and 16 attributes
[11]	DQ on the web	5 dimensions
[24]	Web sites	5 categories and 10 sub-categories
[25]	Organizational networks	6 stages to DQ analysis with several dimensions associated with each one
[26]	Data integration	2 factors and 4 metrics
[27]	Web information portals	2 dimensions

3 Relationships Between the Components of the Model

Based on the previous background, we will determine the relationship between the web portal functionalities and the quality expectations of data consumers. Then, we will present the definition of each function according to [13] and we will show their relationships (see figure 2).

- *Data Points and Integration.* They provide the ability to access information from a wide range of internal and external information sources and display the resulting information at the single point-of-access desktop. The expectations applied to this functionality are: *Content* (Consumers need a description of portal areas covered, use of published data, etc.), *Presentation* (formats, language, and others are very important for easy interpretation) and *Improvement* (users want to participate with their opinions in the portal improvements knowing the result of applying them).
- *Taxonomy.* It provides information context (including the organization-specific categories that reflect and support organization's business), we consider that the expectations of data consumer are: *Content* (consumers need a description of which data are published and how they should be used, easy-to-understand definitions of every important term, etc.), *Presentation* (formats and language in the taxonomy are very important for easy interpretation, users should expect to find instructions when reading the data), and *Improvement* (user should expect to convey his/her comments on data in the taxonomy and know the result of improvements).
- *Search Capabilities.* It provides several services for web portal users and needs searches across the enterprise, World Wide Web, and search engine catalogs and

indexes. The expectations applied to this functionality are: *Quality of values* (Data consumer should expect that the result of searches is correct, current and complete), *Presentation* (formats and language are important for consumers, for the search and for easy interpretation of results) and *Improvement* (consumer should expect to convey his/her comments on data in the taxonomy and know the result of improvements).

- *Help Features*. They provide help when using the web portal. The expectations applied to this functionality are: *Presentation* (formats, language, and others are very important for easy interpretation of help texts) and *Commitment* (consumer should be easily able to ask and obtain answer to any question regarding the proper use or meaning of data, update schedules, etc.).
- *Content Management*. This function supports content creation, authorization, and inclusion in (or exclusion from) web portal collections. The expectations applied to this functionality are: *Privacy* (it should exist privacy policy for all consumers to manage, to access sources and to guarantee web portals data), *Content* (consumers need a description of data collections, that all data needed for an intended use are provided, etc.), *Quality of values* (consumer should expect that all data values are correct, current and complete, unless otherwise stated), *Presentation* (formats and language should be appropriate for easy interpretation), *Improvement* (consumer should expect to convey his/her comments on contents and their management and know the result of the improvements) and *Commitment* (consumer should be easily able to ask and have any question regarding the proper use or meaning of data, update schedules, etc. answered).
- *Process and Action*. This function enables the web portal user to initiate and participate in a business process of portal owner. The expectations applied to this functionality are: *Privacy* (Data consumer should expect that there is a privacy policy to manage the data about the business on the portal), *Content* (Consumers should expect to find descriptions about the data published for the processes and actions, appropriate and inappropriate uses, that all data needed for the process and actions are provided, etc.), *Quality of values* (that all data associated to this function are correct, current and complete, unless otherwise stated), *Presentation* (formats, language, and others are very important for properly interpret data), *Improvement* (consumer should expect to convey his/her comments on contents and their management and know the result of improvements) and *Commitment* (consumer should be easily able to ask and to obtain answer to any questions regarding the proper use or meaning of data in a process or action, etc.).
- *Collaboration and Communication*. This function facilitates discussion, locating innovative ideas, and recognizing resourceful solutions. The expectations applied to this functionality are: *Privacy* (consumer should expect privacy policy for all consumers that participate in activities of this function), and *Commitment* (consumer should be easily able to ask and have any questions regarding the proper use or meaning of data for the collaboration and/or communication, etc).
- *Personalization*. This is a critical component to create a working environment that is organized and configured specifically to each user. The expectations applied to this functionality are: *Privacy* (consumer should expect privacy and security about their personalization data, profile, etc.), and *Quality of values* (data about user profile should be correct, current).
- *Presentation*. It provides both the knowledge desktop and the visual experience to the web portal user that encapsulates all of the portal's functionality. The expectations

applied to this functionality are: *Content* (the presentation of a web portal should include data about covered areas , appropriate and inappropriate uses, definitions, information about the sources, etc.), *Quality of values* (the data of this function should be correct, current and complete.), *Presentation* (formats, language, and others are very important for easy interpretation and appropriate use of portals data.) and *Improvement* (consumer should expect to convey his/her comments on contents and their management and know the result of the improvements).

- *Administration*. This function provides service for deploying maintenance activities or tasks associated with the web portal system. The expectations applied to this functionality are: *Privacy* (Data consumers need security for data about the portal administration) and *Quality of values* (Data about tasks or activities of administration should be correct and complete).
- *Security*. It provides a description of the levels of access that each user or groups of users are allowed for each portal application and software function included in the web portal. The expectations applied to this functionality are: *Privacy* (consumer need privacy policy about the data of the levels of access of data consumers.), *Quality of values* (data about the levels of access should be correct and current.) and *Presentation* (data about security should be in format and language for easy interpretation).

		Web Portal Functionalities											
		Date Probe and Integration	Taxonomy	Search Capabilities	Help Features	Content Management	Process and Action	Collaboration and Communication	Personalization	Presentation	Administration	Security	
Category of Data Consumer Expectations	Privacy					√	√	√	√	√	√	√	√
	Content	√	√			√	√			√	√	√	√
	Quality of Values	√	√	√		√	√		√	√	√	√	√
	Presentation	√	√	√	√	√	√	√	√	√	√	√	√
	Improvement	√	√	√		√	√			√	√	√	√
	Commitment			√	√	√							

Fig. 2. Matrix stating the relationships between data consumer expectations and web portal functionalities

Concerning the relationships established in the matrix of figure 2, we can remark that *Presentation* is the category of data consumer expectation with more relations. This perfectly fits with the main goal of any web applications, which is to be useful and user-friendly for any kind of user.

The next step is to fill in each cell of the matrix with Web DQ/IQ attributes obtained from the study presented in 2.3. As a result of this, we have a subset of DQ/IQ attributes that can be used in a web portal to evaluate data quality. In table 2, we will show the most relevant attributes for each category of data consumer expectations.

To validate and complete this assignation we plan to work with portal data consumers through surveys and questionnaires. Once the validation is finished, we will reorganize the attributes obtaining the final version of the DQ/IQ web portal model.

Table 2. Web Data Quality attributes applied to web portal functionalities in each category

Category of Data Consumer Expectations	Web portal functionalities related to each category
Web DQ/IQ attributes applying almost one functionality in each category	
Privacy	Content management, Process and actions, Collaboration and Communication, Personalization, Administration, Security
Security	
Content	Data Points and Integration, Taxonomy, Content management, Process and actions, Presentation
Accessibility, Currency, Amount of data, Understandability, Relevance, Concise Representation, Validity, Traceability, Completeness, Reliability, Credibility, Timeliness, Availability, Documentation, Specialization, Interpretability, Easy to use	
Quality of data	Data Points and Integration, Search Capabilities, Content management, Process and actions, Personalization, Presentation, Security
Accessibility, Currency, Amount of data, Credibility, Understandability, Accuracy, Expiration, Novelty, Relevance, Validity, Concise Representation, Completeness, Reliability, Availability, Documentation, Duplicity, Specialization, Interpretability, Objectivity, Relevance, Reputation, Traceability, Utility, Value-added, Easy to use	
Presentation	Data Points and Integration, Taxonomy, Search Capabilities, Help Features, Content management, Process and actions, Collaboration and Communication, Presentation, Administration, Security
Amount of data, Completeness, Understandability, Easy to use, Concise Representation, Consistent Representation, Validity, Relevance, Interpretability, User support, Availability, Specialization, Flexibility	
Improvement	Data Points and Integration, Taxonomy, Search Capabilities, Content management, Process and actions, Presentation
Accessibility, Reliability, Credibility, Understandability, User support, Traceability	
Commitment	Help Features, Content management, Process and actions
Accessibility, Reliability, User support,	

4 Validation of the Model

In order to valid our model we plan to elaborate a survey to check the DQ/IQ attributes identified as relevant to the web portals. We will use the *Principles of Survey Research* proposed in [28] where is said that a survey is part of a larger process and recognize that it is not just the instrument for gathering information. In this work the authors identify ten activities in the survey process.

At this moment we are developing the first activities in our survey process. In particular *setting specific and measurable objectives* (in our case this phase consists in

checking the DQ/IQ attributes identified as relevant to the web portals and in obtaining other than were not considered), *planning and scheduling the survey*, *ensuring that appropriate resources are available* and *designing the data collection instrument*.

As survey design we have selected the descriptive design because we try to describe a phenomenon of interest [29] (in our case is to describe the DQ/IQ attributes more relevant for web portal data consumers). We plan to make a questionnaire for each one of the web portal functionalities presented previously. As it is quite impossible to survey the entire population [29], we are developing a web application to be linked in a web portal (www.castillalamancha.es). In that way, the users connected to this portal will be invited to answer some questions (selected randomly between the eleven questionnaires). So, each questionnaire will be constructed for each subject of the survey with the aim of having a correct distribution in the amount of answers given to each question.

The application will have three modules: an administrator module (through it the researcher can generate the questionnaires deciding the number of questions, the type of answer, etc.), an analyzer module (that shows the results: statistics, graphics, ranking of responses, etc.), and a gather module (that presents the questions to the users). So, we will ask each subject about general demographic questions (as the expertise in the use of portals, expertise in technologies, range of age, sex, etc.) together with thirty questions selected randomly from all questions in the eleven questionnaires. When we have enough responses for each question in our questionnaires we will analyze the responses for obtaining a minimum and necessary set of DQ/IQ attributes for each aspect of our model. This set of attributes will be used in order to elaborate a complete framework for evaluating the DQ/IQ of a web portal. For example, we plan to give the minimum value necessary for each attribute for assuring the DQ/IQ quality. If this value is not achieved for some of the attributes, the framework will give some corrective actions applicable in order to have the correct level of quality.

5 Conclusions

The great majority of works found in the literature show that data quality or information quality is very dependent on the context. The increase of the interest in the development of web applications has implied either the appearance of new proposals of frameworks, methodologies and evaluation methods of DQ/IQ or the adaptation of the already-existing ones from other contexts. However, in the web portal context, data quality frameworks do not exist.

In this paper, we have presented a proposal that combines three aspects: (1) a set of web data quality attributes resulting from a data quality literature survey that can be applicable and useful for a web portal, (2) the data quality expectations of data consumer on the Internet, and (3) the basic functionalities for a web portal. These aspects have been related by obtaining a first set of data quality attributes for the different data consumer expectations X functionalities.

Our future work, now in progress, consists of validating and refining this model. First of all, it is necessary to check these DQ/IQ attributes with data consumers in a web portal, for this we plan to make a survey as was presented in the previous section.

Then, once we have validated the model, we will define a framework including the necessary elements to evaluate a DQ/IQ in a web portal. Our aim is to obtain a flexible framework where the data consumer can select the attributes used to evaluate the quality of data in a web portal, depending on the existing functionalities and their personal data quality expectations.

Acknowledgements

This research is part of the following projects: CALIPO (TIC2003-07804-C05-03) supported by the Dirección General de Investigación of the Ministerio de Ciencia y Tecnología (Spain) and DIMENSIONS (PBC-05-012-1) supported by FEDER and by the “Consejería de Educación y Ciencia, Junta de Comunidades de Castilla-La Mancha” (Spain) and CALIPSO (TIN2005-24055-E).

References

1. Strong, D., Y. Lee, and R. Wang, *Data Quality in Context*. Communications of the ACM, 1997. Vol. 40, N° 5: p. 103 -110.
2. Cappiello, C., C. Francalanci, and B. Pernici. *Data quality assessment from the user's perspective*. in *International Workshop on Information Quality in Information Systems, (IQIS2004)*. 2004. Paris, Francia: ACM.
3. Lee, Y., *AIMQ: a methodology for information quality assessment*. Information and Management. Elsevier Science, 2002: p. 133-146.
4. Winkler, W., *Methods for evaluating and creating data quality*. Information Systems, 2004. N° 29: p. 531-550.
5. Marchetti, C., et al. *Enabling Data Quality Notification in Cooperative Information Systems through a Web-service based Architecture*. in *Proceeding of the Fourth International Conference on Web Information Systems Engineering*. 2003.
6. Fugini, M., et al., *Data Quality in Cooperative Web Information Systems*. 2002.
7. Zhu, Y. and A. Buchmann. *Evaluating and Selecting Web Sources as external Information Resources of a Data Warehouse*. in *Proceeding of the 3rd International Conference on Web Information Systems Engineering*. 2002.
8. Bouzeghoub, M. and Z. Kedad, *Quality in Data Warehousing*, in *Information and Database Quality*, M. Piattini, C. Calero, and M. Genero, Editors. 2001, Kluwer Academic Publishers.
9. Katerattanakul, P. and K. Siau, *Information quality in internet commerce desing*, in *Information and Database Quality*, M. Piattini, C. Calero, and M. Genero, Editors. 2001, Kluwer Academic Publishers.
10. Aboelmegeed, M. *A Soft System Perspective on Information Quality in Electronic Commerce*. in *Proceeding of the Fifth Conference on Information Quality*. 2000.
11. Gertz, M., et al., *Report on the Dagstuhl Seminar "Data Quality on the Web"*. SIGMOD Record, 2004. vol. 33, N° 1: p. 127-132.
12. Mahdavi, M., J. Shepherd, and B. Benatallah. *A Collaborative Approach for Caching Dynamic Data in Portal Applications*. in *Proceedings of the fifteenth conference on Australian database*. 2004.
13. Collins, H., *Corporate Portal Definition and Features*. 2001: AMACOM.

14. Kopcsó, D., L. Pipino, and W. Rybolt. *The Assessment of Web Site Quality*. in *Proceeding of the Fifth International Conference on Information Quality*. 2000.
15. Burgess, M., N. Fiddian, and W. Gray. *Quality Measures and The Information Consumer*. in *IQ2004*. 2004.
16. Wang, R. and D. Strong. *Beyond Accuracy: What Data Quality Means to Data Consumer*. *Journal of Management Information Systems*, 1996. 12(4): p. 5-33.
17. Redman, T., *Data Quality: The Field Guide*. 2001: Digital Press.
18. Kerr, K. and T. Norris. *The Development of a Healthcare Data Quality Framework and Strategy*. in *IQ2004*. 2004.
19. Katerattanakul, P. and K. Siau. *Measuring Information Quality of Web Sites: Development of an Instrument*. in *Proceeding of the 20th International Conference on Information System*. 1999.
20. Naumann, F. and C. Rolker. *Assesment Methods for Information Quality Criteria*. in *Proceeding of the Fifth International Conference on Information Quality*. 2000.
21. Pernici, B. and M. Scannapieco. *Data Quality in Web Information Systems*. in *Proceeding of the 21st International Conference on Conceptual Modeling*. 2002.
22. Graefe, G. *Incredible Information on the Internet: Biased Information Provision and a Lack of Credibility as a Cause of Insufficient Information Quality*. in *Proceeding of the Eighth International Conference on Information Quality*. 2003.
23. Eppler, M., R. Algesheimer, and M. Dimpfel. *Quality Criteria of Content-Driven Websites and Their Influence on Customer Satisfaction and Loyalty: An Empirical Test of an Information Quality Framework*. in *Proceeding of the Eighth International Conference on Information Quality*. 2003.
24. Moustakis, V., et al. *Website Quality Assesment Criteria*. in *Proceeding of the Ninth International Conference on Information Quality*. 2004.
25. Melkas, H. *Analyzing Information Quality in Virtual service Networks with Qualitative Interview Data*. in *Proceeding of the Ninth International Conference on Information Quality*. 2004.
26. Bouzeghoub, M. and V. Peralta. *A Framework for Analysis of data Freshness*. in *IQIS2004*. 2004. Paris, France: ACM.
27. Yang, Z., et al., *Development and validation of an instrument to measure user perceived service quality of information presenting Web portals*. *Information and Management*. Elsevier Science, 2004. 42: p. 575-589.
28. Pfleeger, S. and B. Kitchenham, *Principles of Survey Research Part1: Turning Lemons into Lemonade*. *Software Engineering Notes*, 2001. 26(6): p. 16-18.
29. Pfleeger, S. and B. Kitchenham, *Principles of Survey Research Part2: Designing a Survey*. *Software Engineering Notes*, 2002. 27(1): p. 18-20.

Design for Environment-Friendly Product

Hak-Soo Mok¹, Jong-Rae Cho², and Kwang-Sup Moon¹

¹ Department of Industrial Engineering Pusan National University,
Busan 609-735, Korea

hsmok@pusan.ac.kr, ksmoon@dreamwiz.com

² R&D Division for Hyundai Motor Company & Kia Motors Corporation,
Gyunggi-Do 445-706, Korea
muyoung@hyundai-motor.com

Abstract. This paper presents an approach for systematically integrating design for x (DFX) into environmentally conscious product design. Evaluation methods and algorithms for design for assembly, design for disassembly, and design for recycling are proposed to estimate the assembly times and the disassembly times. By these results could be evaluated the assemblability, disassemblability, and recyclability of parts. Finally, a new integrated DFX system is implemented to help designers of product analyze environment-friendly products during the design process.

1 Introduction

In recent years, environmental issues have been becoming increasingly important to product designers and manufacturers (Holloway, 1998). Thus, environmentally conscious design (Eco-design) methodologies should be developed and integrated for a product's entire life cycle. Generally, Eco-design includes a life cycle assessment (LCA) and design for x (DFX) including design for assembly (DFA), design for disassembly (DFD), and design for recycling (DFR). Figure 1 outlines the research and development methodologies of Eco-design. In the DFA and DFD modules, new methodologies are presented using process data and geometrical data. To evaluate the recyclability of parts, the outputs of DFD and LCA and a cost analysis are used in the DFR module.

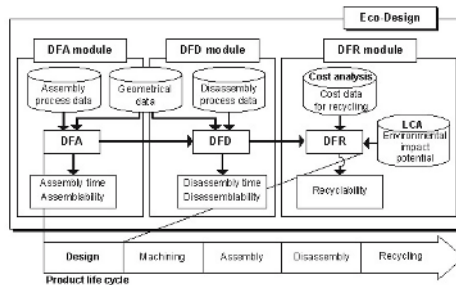


Fig. 1. Scope of Research

2 Design for Assembly and Disassembly

Assembly and disassembly are necessary and critical processes for the end-of-life (EOL) of a product, because they have an effect on recycling operations (Redford., 1994). This paper proposes a new DFA methodology that can estimate the standard assembly time and evaluate the assemblability of parts. And, to facilitate their disassembly of parts for recycling, this study attempts to develop a design for disassembly evaluation metrics to be used when designing new products.

2.1 Procedures for DFA and DFD

This study suggests three main decision factors - handling, access and insertion - as the criteria for the analysis of ssembly processes. The influencing factors associated with these three decision factors are determined in order to estimate assembly time and to evaluate assemblability. The decision factors for the disassembly process are fixing, grasping, access, disassembly, and handling. Finally, the total assembly and disassembly time and assemblability and disassemblability are estimated by the sum of the time and the disassemblability scores of the decision factors. The overall procedure for DFA and DFD is shown in Figure 2.

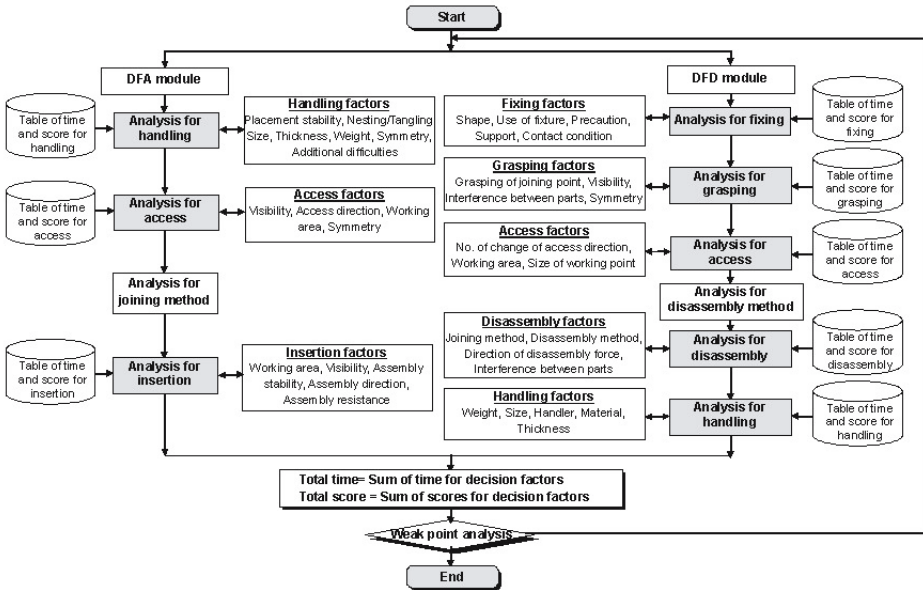


Fig. 2. Procedures for DFA and DFD

2.2 Time and Score Tables for Decision Factors

Table 1 shows an example of the decision and influencing factors and their respective levels for assembly. To obtain tables that include the time and score of each decision factor for assembly or disassembly, the influencing factors are divided into levels. And then the weights numbers in the brackets with the decision and the influencing factors are calculated by the AHP (Analytic Hierarchy Process) (Saaty, 1986), and difficulty scores are given to each level. This paper subdivided the difficulty scores into 5 sets: 1 (Good), 3 (Not bad), 5 (Not good), 7 (Bad), and 9 (Absolutely bad). The 1 to 9 scaling has proven to be an acceptable scale and is recommended for use in the AHP (Harker et al., 1987). The times for assembly and disassembly are estimated through a motion analysis of the assembly and disassembly operations. This paper mixed MTM (Methods Time Measurement) and the WF (Work factor) in the motion analysis. MTM has many distinct advantages, but because it is not sufficient to describe all of the assembly and disassembly processes, we revised it according to the WF.

Table 1. Decision and Influencing Factors and Their Levels for Assembly

Criteria	Influencing factors		Level		
			1	2	3
Handling (0.12)	Placement stability (0.05)		Good (1)	Bad (5)	
	Nesting/Tangling (0.16)		No (1)	Yes (7)	
	Size (0.22)	I, II	$1 < S \leq 20$ (1)	$20 < S \leq 50$ (3)	$0 < S \leq 1, S > 50$ (9)
		III	$S \leq 100$ (1)	$S > 100$ (7)	
	Thickness (0.07)		$0 < T \leq 5$ (1)	$5 < T \leq 10$ (5)	
	Weight (0.09)	I, II	$W \leq 0.3$ (1)	$0.3 < W \leq 2$ (7)	
		III	$W \leq 3$ (1)	$W > 3$ (9)	
	Symmetry (0.12)	I	$\alpha + \beta \leq 360$ (1)	$360 < \alpha + \beta \leq 540$ (3)	$540 < \alpha + \beta \leq 720$ (5)
II		$A \leq 180$ (1)	$180 < \alpha \leq 360$ (5)		
III		$\alpha \leq 180$ (1)	$\alpha = 180$ (5)		
Additional difficulties (0.29)		No (1)	Yes (7)		
Access (0.25)	Visibility (0.30)		Good (1)	Bad (7)	
	Assess direction (0.09)		Vertical(Gravity) (1)	Horizontal(oblique) (5)	Combined (7)
	Working area (0.55)		Good (1)	Bad (7)	
	Symmetry (0.06)		$\beta = 0, \beta = 90$ (1)	$\beta = 180, \beta = 360$ (5)	
Insertion (0.63)	Bolting	Working area	Good (1)	Bad (7)	
		Visibility	Good (1)	Bad (7)	
		Assembly stability	Good (1)	Bad (7)	
		Assembly direction	Vertical(Gravity) (1)	Horizontal(oblique) (5)	Combined (9)
		Assembly resistance	No(1)	Yes(5)	

Figure 3 shows the process for obtaining the assembly or disassembly time table, including the influencing factors, through motion analysis. When the levels of factor 4 and factor 5 are changed to 4' and 5", the assembly or disassembly time is extended by penalty times 14 and 30 TMU, respectively. The relationship between the penalty time and the change of the level of factor can be presented as

$$T_{1,2,3,4,5} = T_{1,2,3,4,5} + PT_{1,2,3,4 \rightarrow 4',5} + PT_{1,2,3,4,5 \rightarrow 5} \tag{1}$$

$$= T_{1,2,3,4,5} + PT_{1,2,3,4 \rightarrow 4',5} + PT_{1,2,3,4,5 \rightarrow 5} + \alpha$$

where $T_{1,2,3,4,5}$ is the assembly or disassembly time when influencing factors 1, 2, 3, 4 and 5 are changed to 1, 2, 3, 4', and 5'', and $PT_{1,2,3,4 \rightarrow 4',5}$ represents the penalty time when influencing factor 4 is changed to 4'. α denotes the added

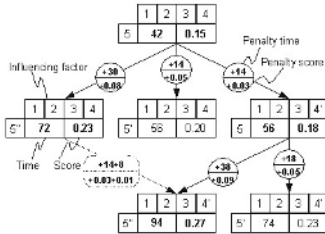


Fig. 3. Calculation Process for Time and Scores in Tables

Table 2. Time and Score Tables of Tying, Painting and Snapping

Time and score table for tying	Material of tying: Flexible (Rubber...)				Material of tying: Rigid (Steel...)			
	Type of tying: Simple		Type of tying: Complex		Type of tying: Simple		Type of tying: Complex	
	Fixing: Yes	Fixing: No (Support)	Fixing: Yes	Fixing: No (Support)	Fixing: Yes	Fixing: No (Support)	Fixing: Yes	Fixing: No (Support)
	0	1	2	3	4	5	6	7
Working area: Good	110	148	264	300	166	198	334	390
	0.63	0.78	1.99	2.14	1.54	1.69	2.90	3.05
Working area: Bad	144	192	342	376	210	252	416	480
	1.69	1.84	3.05	3.20	2.60	2.75	3.96	4.11

Time and score table for painting	Precaution for other part: Not needed				Precaution for other part: Needed			
	Hole, Slot: Not existed		Hole, Slot: Existed		Hole, Slot: Not existed		Hole, Slot: Existed	
	Shape: Simple	Shape: Complex	Shape: Simple	Shape: Complex	Shape: Simple	Shape: Complex	Shape: Simple	Shape: Complex
	0	1	2	3	4	5	6	7
Working area: Good	180	220	246	270	264	294	316	348
	0.63	1.23	1.27	1.88	1.51	2.12	2.15	2.76
Working area: Bad	258	300	348	386	380	422	480	548
	1.23	1.84	1.88	2.48	2.12	2.72	2.76	3.36

Time and score table for Snapping	Need other force to other part for fitting: No				Need other force to other part for fitting: Yes			
	Orientation, Alignment: Good		Orientation, Alignment: Bad		Orientation, Alignment: Good		Orientation, Alignment: Bad	
	Resistance		Resistance		Resistance		Resistance	
	Small	Large	Small	Large	Small	Large	Small	Large
0	1	2	3	4	5	6	7	
Working area: Good	40	52	72	90	60	78	96	116
	0.63	1.21	2.14	2.72	1.01	1.59	2.52	3.10
Working area: Bad	60	74	98	120	82	104	130	156
	1.18	1.76	2.70	3.28	1.56	2.14	3.07	3.65

penalty time due to the changing levels. In Fig. 3, α is 8 TMU. The scores of tables can be calculated by equation (2).

$$S = \sum_{i=1}^n \sum_{j=1}^m w_i w_j s_j \quad (2)$$

where w_i and w_j represent the weights of the decision and influencing factors I and j respectively, and s_j is the difficulty score of influencing factor j . Table 2 shows an example of time and score tables for the assembly method: tying, painting, and snapping. Time and score tables can be obtained for each decision factor of assembly and disassembly. However, because the influencing factors are different according to the assembly and disassembly methods, the time and score tables for decision factor, insertion and disassembly are obtained for each assembly and disassembly method.

3 Design for Recycling

It is currently widely acknowledged that the most ecologically sound way to treat a worn-out product is recycling (Mildenberger, 2000). This paper proposes a systematic approach in which a product is designed to evaluate its recyclability. All DFX areas including DFA, DFD and LCA are considered, and a methodology is suggested for integrating these areas with DFR.

3.1 LCA and Cost Analysis

LCA is a methodology for evaluating the environmental effects occurring throughout the entire life cycle of a product, process, or activity (Krikke, 1999). In this environmental support system, environmental information such as the environmental potential of materials and environmental impact, is given over to the LCA stage of the assembly and disassembly processes. Through the LCA, environmental impacts such as resource depletion, global warming, ozone depletion, photochemical oxidation creation, acidification, and others, are assessed. Generally, a cost analysis is achieved by a comparison of cost and profit. In this paper, the cost analysis for recycling is defined as

$$C_R = C_C - C_D - C_M + C \quad (3)$$

where C_R represents the recycling cost, and C_C , C_D , C_M , and C_S are the collection, disassembly, maintenance, and selling costs, respectively.

3.2 Evaluation of Recyclability of Parts

In this paper, the recyclability of parts is computed based on equation (4), as follows.

$$S_{\text{Recyclability}} = aS_{\text{Disassemblability}} + bS_{\text{LCA}} + cS_{\text{Cost}} \quad (4)$$

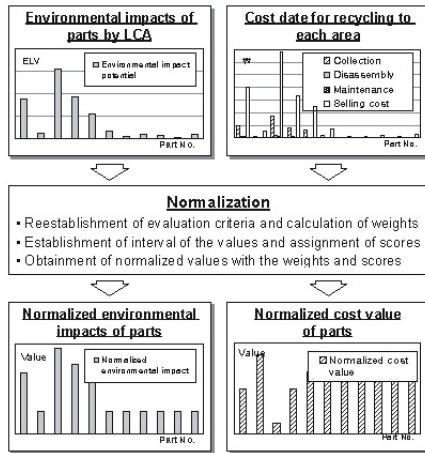


Fig. 4. Normalization Process for Integration

$S_{Disassemblability}$ represents the disassemblability score already evaluated in the DFD module, S_{LCA} is the score of the environmental impact potential of parts computed by the LCA, and S_{Cost} is the score determined by the cost analysis for recycling. The values a , b , and c are the weights that will be computed by the AHP. The recyclability can be considered as an integrated evaluation of technical information such as disassemblability, environmental information such as environmental impacts by LCA, and economical information such as cost data by cost analysis.

The environmental impacts and cost values of each part should be normalized to obtain the final recyclability of parts because these are different from disassemblability. Figure 4 shows the normalizing process of environmental impacts and cost data. The criteria that are used to obtain the environmental impacts such as resource depletion, global warming, ozone depletion, acidification, and others, are re-established and their weights are obtained. The values of these environmental impacts are divided into levels and scores are assigned to each level. Eventually, the weights and scores result in the normalized environmental impact values.

4 Development of DFX System

There is a need for such a DFX system that enables designers to effectively analyze the ease of assembly, disassembly, and recycling of the products. This paper presents a new approach to DFA, DFD, and DFR to help designers to estimate the assembly and disassembly times and to evaluate the assemblability, disassemblability, and recyclability.

4.1 DFA and DFD Systems

Because the assembly time and assemblability are calculated by the three decision factors as the criteria for the evaluation of assembly, the DFA system needs three main input forms: handling, access, and insertion (see Figure 5).

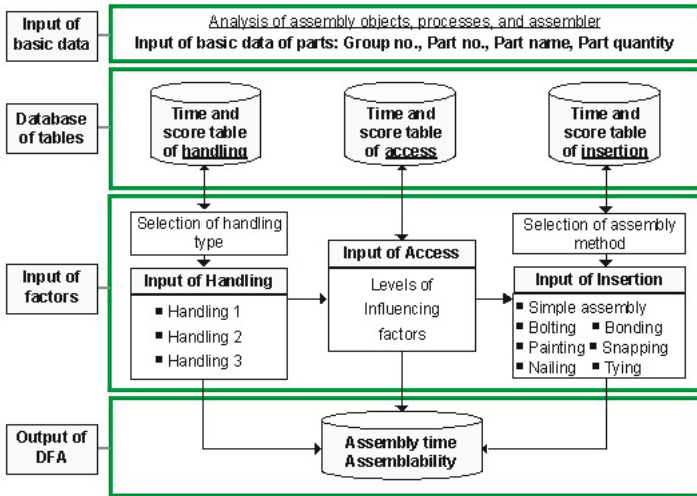


Fig. 5. System Concept for DFA

The DFD system is similar to the DFA system. However, five input processes - fixing, grasping, access, disassembly, and handling - are analyzed to get the disassembly time and disassemblability of parts. In the DFA and DFD modules, there are four output forms: assembly time, assemblability, disassembly time, and disassemblability. Figure 6 shows the output forms of assembly time and disassemblability and the score, of parts. Figure 6 shows the input forms of decision factor insertion and disassembly used to determine the assembly and disassembly time and the assemblability and disassemblability. As mentioned before, there are three and five decision factors with their influencing factors and levels. In the case of insertion, a new frame that includes the influencing factors is shown when the assembly method is chosen.

4.2 DFR System

The developed DFR system guides its users through three main input modules, LCA, DFD, and cost analysis.

The results for the disassemblability of parts that are obtained in the DFD module, the environmental impacts by LCA, and the cost values by cost analysis for recycling are normalized to obtain the recyclability of parts (see Figure 7). The output form of the recyclability of parts is shown in Figure 8. The normalized values of environmental impact, cost, and disassemblability are given in a table.

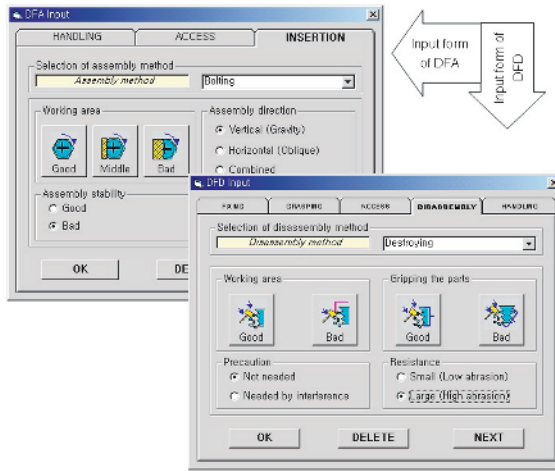


Fig. 6. Input Forms of DFA and DFD

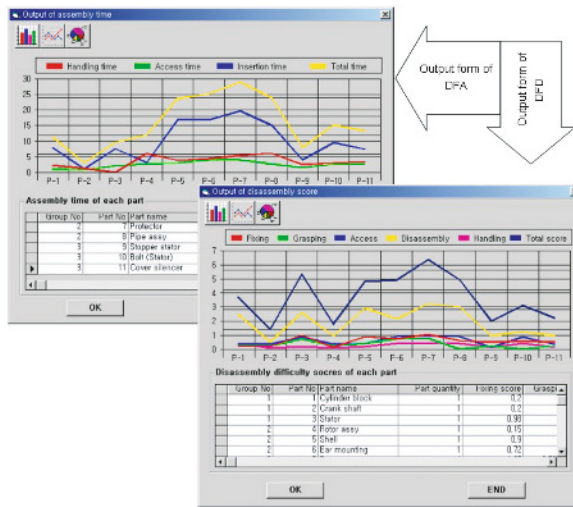


Fig. 7. Output Forms for Assembly Time and Disassembly Difficulty Score

5 Conclusions

This paper has demonstrated a new approach for DFA, DFD, and DFR, and has shown how these DFX methodologies can be integrated into environmentally conscious design. This paper has also presented a new integrated DFX system to help designers to estimate the assembly and disassembly times, and to evaluate the assemblability, disassemblability, and recyclability. Following the

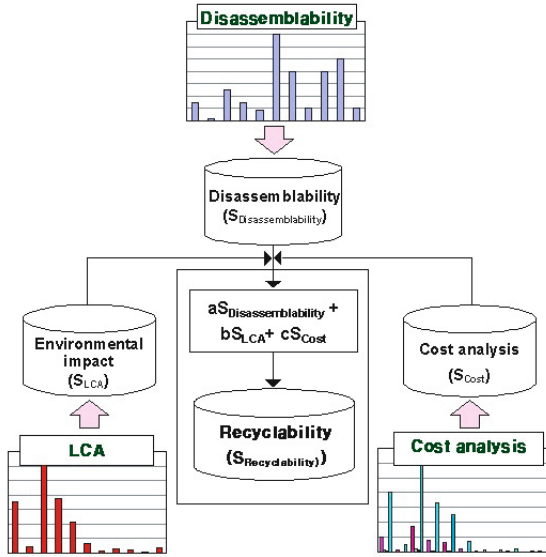


Fig. 8. System Concept for DFR

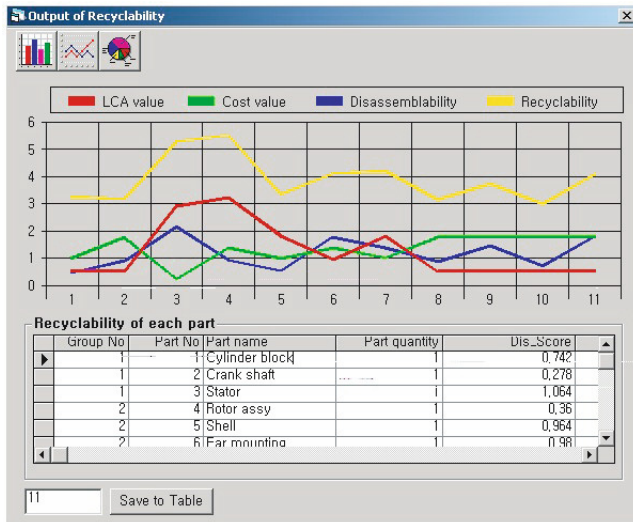


Fig. 9. Output Form for Recyclability

above approach, companies will be able to design products that have low environmental impacts. Our future work will focus on generating redesign alternatives for environmentally friendly products.

Acknowledgment. This work was supported by Cleaner Production Technology Development of Ministry of Commerce, Industry and Energy and Brain Busan 21 Project in 2004.

References

- 1 Harker, P. et al., "The Theory of Ratio Scale Estimation: Saaty's Analytic Hierarchy Process", *Management Science*, Vol. 33, pp.1383-1402, 1987.
- 2 Holloway, L., "Materials Selection for Optimal Environmental Impact in Mechanical Design", *Materials and Design*, Vol. 19, pp.133-143, 1998.
- 3 Krikke, H. et al., "Business Case Roteb: Recovery Strategies for Monitors", *Computers & Industrial Engineering*, Vol. 36, pp.739-757, 1999.
- 4 Mildenberger, U., and Khare, A., "Planning for an Environment-Friendly Car, *Tech-novation*", Vol. 20, pp.201-214, 2000.
- 5 Redford, A. et al., *Design for Assembly*, McGraw-Hill, Inc., 1994.
- 6 Saaty, T., "Axiomatic Foundation of the Analytic Hierarchy Process", *Management Science*, Vol. 32, pp.841-855, 1986.

Performance of HECC Coprocessors Using Inversion-Free Formulae*

Thomas Wollinger¹, Guido Bertoni², Luca Breveglieri³, and Christof Paar⁴

¹ Escript GmbH - Embedded Security - Bochum, Germany
twollinger@escript.com

² STMicroelectronics - Advanced System Technology - Agrate B., Milano, Italy
guido.bertoni@st.com

³ Politecnico di Milano, Italy
breveglieri@elet.polimi.it

⁴ Communication Security Group (COSY),
Ruhr-Universitaet Bochum, Germany
cpaar@crypto.rub.de

Abstract. The HyperElliptic Curve Cryptosystem (HECC) was quite extensively studied during the recent years. In the open literature one can find results on how to improve the group operations of HECC as well as the implementations for various types of processors. There have also been some efforts to implement HECC on hardware devices, like for instance FPGAs. Only one of these works, however, deals with the inversion-free formulae to compute the group operations of HECC.

We present inversion-free group operations for the HEC $y^2 + xy = x^5 + f_1x + f_0$ and we target characteristic-two fields. The reason is that of allowing a fair comparison with hardware architectures using the affine case presented in [BBWP04]. In the main part of the paper we use these results to investigate various hardware architectures for a HECC VLSI coprocessor. If area constraints are not considered, scalar multiplication can be performed in 19,769 clock cycles using three field multipliers (of type $D = 32$), one field adder and one field squarer, where D indicates the digit-size of the multiplier. However, the optimal solution in terms of latency and area uses two multipliers (of type $D = 4$), one addition and one squaring. The main finding of the present contribution is that coprocessors based on the inversion-free formulae should be preferred compared to those using group operations containing inversion. This holds despite the fact that one field inversion in the affine HECC group operation is traded by up to 24 field multiplications in the inversion-free case.

Keywords: Hyperelliptic curve cryptosystem, inversion-free formulae, hardware architecture, high performance explicit formulae, VLSI coprocessor.

1 Introduction

Koblitz and Miller proposed, independently from each other, Elliptic Curves (EC) for public-key cryptography. The generalizations of ECs are called

* Accepted on 31 Jan 2005 for ISH 2005, published in 2006.

HyperElliptic Curves (HEC) and were first proposed for cryptographic use in [Kob88]. It is important to point out that hyperelliptic curve cryptosystems are promising because of the *short* operand sizes, compared to other public key schemes. We introduce an analysis of the hardware implementation of the inversion-free HEC group operations, including a comparison with affine hardware architectures. Previously there have been some efforts for implementing a HECC coprocessor on FPGAs; however our contribution is the first one containing a thorough analysis of the different design options and a comparison between the two possible coordinate systems (affine and projective). Our analysis is based on the derived group operations and on the underlying field $\mathbb{F}_{2^{83}}$.

We investigated various hardware architectures by parallelizing the scalar multiplication of HECC at the following three levels: the field operation level, the group operation level and the scalar multiplication level. At the field operation level we used multipliers that handle different numbers of bits in parallel. At the group operation level we used different numbers of field multipliers. Our investigation of the parallelism reachable for the scalar multiplication level was based on the idea of overlapping the computation of consecutive group operations.

The scalar multiplication can be performed most efficiently in 19,769 clock cycles using a coprocessor design providing three field multipliers (of type $D=32$), one field adder and one field squarer. In order to find the optimal design for the HECC coprocessor, we considered speed as well as (silicon) area, namely the area-time product. The lowest area-time product was achieved by using two multipliers (of type $D = 4$), one adder and one squarer.

Furthermore, we compared the results of this paper with those published for the group operations using inversion [BBWP04]. Thus we found out that the most efficient way to implement a HECC VLSI coprocessor is to use inversion-free formulae. If we compare the best coprocessor configuration for affine coordinates with that using inversion-free formulae, we see that a) the latency is better up to a factor of almost 3 and b) the area-time product figure for the inversion-free cases is always better. Hence, HECC coprocessors ought to be based on the inversion-free coordinate system. This result could be achieved even by trading one inversion in the affine group operation with up to 24 field multiplications in the inversion-free case.

The paper is organized as follows. Section 2 presents our improved inversion-free HEC group operations and Section 3 describes the architecture of the HECC coprocessor as well as the different levels of parallelization. In Section 4 we present and discuss results. Conclusions list final considerations.

2 Inversion-Free Formulae

In the recent years a considerable effort has been focused on improving the group operations of the hyperelliptic curve cryptosystem (HECC), see [PWP04] for a summary. Analogously to EC, there also exist inversion-free group operations for HECC [MDM⁺02, Lan03]. Not having to use inversions might also be

profitable for HECC, however the overhead in multiplications is very high compared to the simpler ECC case. We investigated the needed field operations of the inversion-free group operations considering a genus-2 HEC $y^2 + h(x)y = f(x)$, with $h(x) = x$ and $f(x) = x^5 + f_1x + f_0$. In addition, the characteristic of the underlying field was fixed to 2. We do not know of any security limitations using this kind of curves. The main reason of using these parameters, was to provide a fair comparison with the affine HECC coprocessor presented in [BBWP04]. Furthermore, we were able to reduce the number of finite field operations, see Table 1.

Table 1. Efficient inversionfree group operations for genus-2 HEC

coordinate system		curve properties	addition	doubling
affine	[Lan03]	$h_2 = 0, h_i \in \mathbb{F}_2$ $f_4 = 0$	I+21M+3S	I+17M+5S
	[?]	$h(x) = x$ $f_4 = f_3 = f_2 = 0$	-	I+9M+6S
projective	[MDM ⁺ 02]	N.A.	67M	42M
	[Lan03]	$h_2 = 0, h_i \in \mathbb{F}_2$ $f_4 = 0$	47M + 4S	40M + 7S
	our work	$h(x) = x$ $f_4 = f_3 = f_2 = 0$	45M + 5S	31M + 6S

3 Architecture of the HECC Coprocessor

Our coprocessor is designed according to a rather standard architecture, consisting of a 3-bus loop scheme that connects a set of functional units and a register file. A control unit drives the various function units and the register file. The register file stores temporary intermediate values and final results. The size of each register is the dimension of the field, namely 83 bits. The register file has two output ports to feed the operands to the function units and one input port to receive the result. The processor allows to load two field elements and store one field element at every clock cycle. This guarantees feasibility and ease of implementation. However, at any given clock cycle only one field operation can start. If the operation is unary, such as inversion, one input bus remains idle.

In Table 2 we give the area and latency for each arithmetic function unit we used. The given estimates assume 2-input gates and optimal field generator polynomials of type $F(x) = x^m + \sum_{i=0}^t f_i x^i$, where $m - t \geq D$. One observes that the gate-consuming function units are the multipliers and the inverter, while in comparison the area used for the adder and the squarer is negligible. In the case of multipliers and inverter the numbers of XOR and AND gates are the same. These two considerations allow us to equal the costs of the XOR and AND gates, for the reason that, when comparing different system configurations, we do not want to use absolute gate sizes, but relative ones.

We have assumed that the different components of the system work at the same frequency. Such an assumption might not be true. Usually, the frequency of this type of cryptosystem is dominated by the multiplier, and a smaller digit-size

Table 2. Components of the coprocessor: Area and time

	Area		Latency [clock cycles]
		Gate count	
Add	$\lceil m \rceil$ XOR	$\lceil m \rceil$	1
Sqr [OP00]	$\lceil 4(m-1) \rceil$ XOR	$\lceil 4(m-1) \rceil$	1
Mul [SP97]	$\lceil D \cdot m \rceil$ AND & $\lceil D \cdot m \rceil$ XOR	$\lceil 2Dm \rceil$	$\lceil m/D \rceil$
Inv [BCH93]	$\lceil 6 \cdot m + \log_2 m \rceil$ AND & $\lceil 6 \cdot m + \log_2 m \rceil$ XOR	$\lceil 2(6m + \log_2 m) \rceil$	$2 \cdot m$

will yield to a higher clock frequency and thus will speed-up the whole system. For a correct estimation of the impact of the different clock frequencies, circuit synthesis is necessary.

In our analysis we considered different levels of parallelization for the HECC coprocessor:

1. Parallelization at field operation level: we achieved it by varying the digit size of the multiplier.
2. Parallelization at group operation level: our architecture of the HEC coprocessor allows to change the number of used multipliers.
3. Parallelization at scalar multiplication level: we achieved it by overlapping two consecutive group operations.

Taking into consideration all the different parallelization levels, we were able to identify the best architecture considering the execution time as well as the area requirements.

We coded a software library that schedules the HECC scalar multiplication mapping it onto the proposed hardware architecture. This software tool applies the so-called Operation Scheduling [Gov03] procedure for parallelizing the sequence of group operations computing the HECC scalar multiplication, and it works according to the As Soon As Possible (ASAP) scheduling policy. We then constrain the scheduling policy by imposing limits on the available hardware resources (i.e. type and number of arithmetic function units) and realistic time delays required to execute the field operations. A more detailed description of the scheduling methodology can be found in [BBWP04].

4 Analysis and Results

The main contribution of this paper is to identify the optimal architecture for the HECC coprocessor. In order to do so we first need to investigate the different architectures for the inversion-free group operations. Afterwards our results are compared to those valid for affine coordinates, as they are presented in [BBWP04].

We target genus-2 HECs defined over a finite field $\mathbb{F}_{2^{83}}$ and we use the derived group operations represented by means of projective coordinates. As for the evaluation of the latency of the scalar multiplication operation (the so-called kD operation), we examined some average cases. Hence the 160-bit long integer k contains the same number of 0s and 1s.

All our results are presented against different digit-sizes, we vary the number of multiplier units and we present the parallelism at the bit level and at the field operation level, respectively. The parallelism at the scalar multiplication level, namely the overlapping between two consecutive group operations, applies to the figures given for the scalar multiplication and the area-time product. We choose not to upper bound the number of used registers. All the considered system configurations require 21 registers for storing temporary values, where each register stores a field element of 83 bits. One could reduce the number of registers, at the cost of some additional latency, by avoiding the overlapping of two consecutive group operations, for example.

4.1 Latency

In this subsection we present our results targeting the latency of the complete scalar multiplication, see Table 3.

Table 3. Latency of the scalar multiplication (in clock cycles, group order $\approx 2^{160}$)

Digit-size	Number of multipliers			
	1	2	3	4
2	356,537	193,176	130,652	98,888
4	181,670	97,355	69,730	55,646
8	98,400	53,243	38,935	31,953
16	56,525	32,835	24,747	22,974
32	31,702	21,209	19,769	20,490

The conclusion one can draw from the table is that with increasing hardware resources (more multiplier units and higher digit size), latency drops. Thus we can compute the scalar multiplication of HECC in a shorter time. Scalar multiplication can be performed most efficiently in 19,769 clock cycles by providing three field multipliers (of type $D = 32$), one field adder and one field squarer (Table 3, bottom row).

Our scheduler is based on the method known as Operation Scheduling (and works according the ASAP scheduling policy), which does not grant an optimal solution [Gov03]; this is evident in the case of 4 multipliers of type $D = 32$ (Table 3, 5-th row).

When carefully inspecting the results, we see that adding more hardware resources might be unnecessary.

4.2 Area-Time Product

An optimal implementation should achieve the maximum throughput and should consume the minimum area (contrary to some traditional cryptographic implementations, where only best time performances were taken into care). Hence we consider both the hardware requirements and the time constraints of the cryptographic application.

Table 4. Normalized area-time product for the scalar multiplication

Digit-size	Number of multipliers			
	1	2	3	4
2	1.2207	1.1024	1.0438	1.0157
4	1.0367	1	1.0346	1.0796
8	1.0107	1.0330	1.1109	1.2034
16	1.0967	1.2367	1.3839	1.7043
32	1.1940	1.5733	2.1885	3.0167

In Table 4 we show the area-time product for the different design options. The analysis exhibits the area-time product normalized with respect to the lowest area-time product. Table 4 shows that the architecture using two multipliers (of type $D = 4$), one adder and one squarer, achieves the best area-time product and therefore is the optimal architecture.

4.3 Comparing Affine and Projective HECC

In this section we put our results into perspective with the HECC VLSI coprocessor implemented in affine coordinates, presented in [BBWP04]¹. This comparison is possible as the used methodology is the same, e.g. identical curve parameters. Figure 1 compares the latencies of the affine and projective HECC coprocessors; the lower bars identify the preferable coprocessor. One can draw the following conclusions:

- In both coordinate systems latency drops by providing additional hardware resources (by increasing the digit-size and the number of multiplier units).
- If the context allows to use projective coordinates, they are always preferable in terms of latency and area. However, affine coordinates might find an appropriate use in low-area applications.
- One very important result in this contribution is that high-speed implementations definitely ought to use projective coordinates, as it can be seen in Figure 1, where projective coordinates using large digit-sizes result in the lowest latency.

¹ Note, that it is not fair to compare our results with the previous work implementing HECC on FPGAs. The reason is the different architecture that is used.

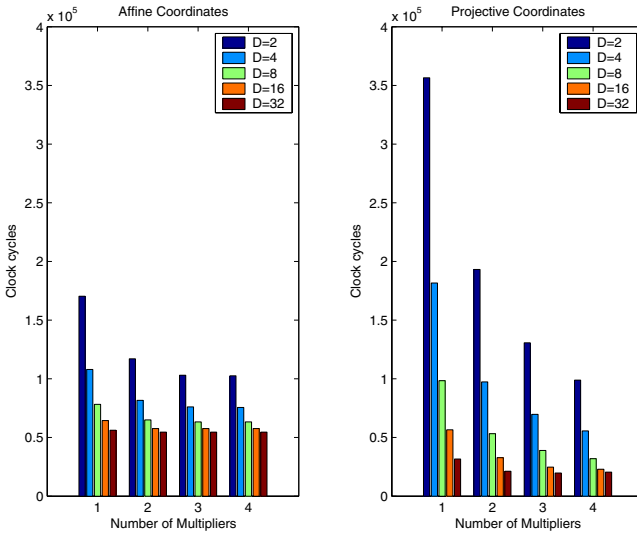


Fig. 1. Latency comparison between affine and projective HECC coprocessors

If comparison between affine and projective coordinates is limited only to the scalar multiplication latency, then the projective processor is better than the affine one. The reason is that the design based on projective coordinates can calculate a scalar multiplication in 19,769 clock cycles, while the affine one can not work faster than in 54,593 clock cycles. This means a speed-up of 2.7, approximately.

In Figure 2 we compare the five best Area-Time (AT) product figures using the two different coordinate systems. For each system configuration the area-time product and the latency are reported. Both these measures are normalized with respect to those of the system configuration with the best AT product. The left-most bars show the figures for affine coordinates and the five right-most bars show those for the projective case. The design option used for the HECC VLSI coprocessor is written at the bottom of each bar (M denotes the multipliers).

It is interesting to analyze the implicit parallelism level achievable by using the projective formulae. By examining the best five area-time products, there appears to exist only one system configuration with one multiplier. In the case of the affine coordinates, we see that the two best solutions contain only one multiplier. Hence, potentially we cannot parallelize as many field operations.

Another general statement one can draw from Figure 2 is that systems using projective coordinates are always better, in terms of the area-time product. We realize that there is a gap of about 35% in the area-time product between the affine and projective coordinates.

Note that projective coordinates are not always better in terms of latency. The reason is the area-time product metric. Hence, the projective systems reaching a lower latency and a better AT product, compared to the affine ones, have

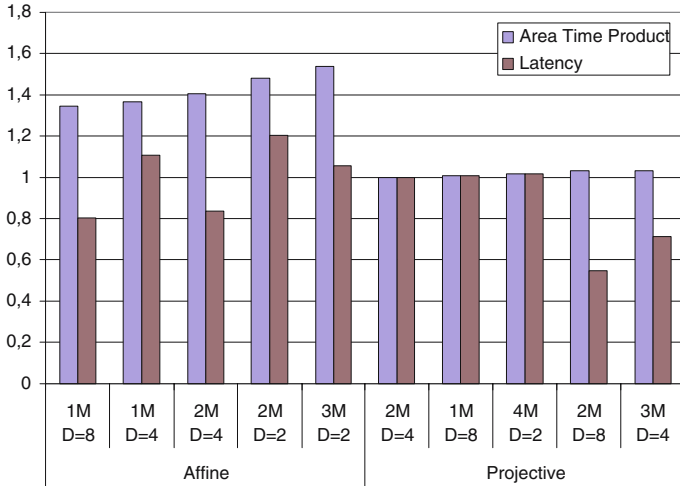


Fig. 2. Comparison of the best five area-time product figures using and an affine and a projective HECC coprocessor

smaller area requirements. However, if we look at the systems with similar area-usage requirements, (i.e., if we compare the projective system using 3M and $D = 4$, with the affine one using 2M and $D = 4$), the AT product is better, as well as the latency and the area of the coprocessor based on projective coordinates. Latency is 17% better than in the affine case.

Hence, the *main* finding of this contribution is that the *projective* coprocessors are more flexible than the affine ones, and thus allow the designer to choose the best compromise in terms of required latency and silicon area. If the application imposes to perform data conversion as well, the affine coprocessors become attractive for low-area requirements.

5 Conclusions

We presented newly derived explicit formulae for HECC using the inversion-free approach. Our formulae are better, up to 22.5%, than the best previous ones, and have been used as the basis of an extensive study of different architecture options for a HECC VLSI coprocessor. We analyzed the parallelization at field operation level, group operation level and scalar multiplication level. The analysis was carried out by using a HEC $y^2 + h(x)y = f(x)$, with $h(x) = x$, $f(x) = x^5 + f_1x + f_0$ and a base field $\mathbb{F}_{2^{83}}$.

Comparing our results based on projective coordinates with those based on affine coordinates for HECC, we can draw the following conclusions: a) considering only the latency the coprocessors based on projective coordinates lead to a higher performance (latency is better up to a factor of almost 3), and b) considering the area-time product the projective coprocessor is preferable as well.

Hence, the group operations computed using inversion-free formulae ought to be the appropriate choice of any HECC VLSI hardware implementation.

References

- [BBWP04] G. Bertoni, L. Breveglieri, T. Wollinger, and C. Paar. Finding optimum parallel coprocessor design for genus 2 hyperelliptic curve cryptosystems. In *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on*, volume 2, pages 538–544. IEEE Computer Society, November 2004.
- [BCH93] H. Brunner, A. Curiger, and M. Hofstetter. On Computing Multiplicative Inverses in $GF(2^m)$. *IEEE Transactions on Computers*, 42:1010–1015, August 1993.
- [Gov03] R. Govindaraian. *Instruction scheduling*. CRC Press, the compiler design handbook edition, 2003.
- [Kob88] N. Koblitz. A Family of Jacobians Suitable for Discrete Log Cryptosystems. In Shafi Goldwasser, editor, *Advances in Cryptology - Crypto '88*, volume 403 of *Lecture Notes in Computer Science*, pages 94 – 99, Berlin, 1988. Springer-Verlag.
- [Lan03] T. Lange. Formulae for Arithmetic on Genus 2 Hyperelliptic Curves, 2003. Available at <http://www.ruhr-uni-bochum.de/itsc/tanja/preprints.html>.
- [MDM⁺02] Y. Miyamoto, H. Doi, K. Matsuo, J. Chao, and S. Tsuji. A Fast Addition Algorithm of Genus Two Hyperelliptic Curve. In *The 2002 Symposium on Cryptography and Information Security — SCIS 2002, IEICE Japan*, pages 497 – 502, 2002. in Japanese.
- [OP00] G. Orlando and C. Paar. A High-Performance Reconfigurable Elliptic Curve Processor for $GF(2^m)$. In Ç. K. Koç and C. Paar, editors, *Cryptographic Hardware and Embedded Systems — CHES 2000*, volume LNCS 1965. Springer-Verlag, 2000.
- [PWP04] J. Pelzl, T. Wollinger, and C. Paar. High performance arithmetic for special hyperelliptic curve cryptosystems of genus two. In *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on*, volume 2, pages 513 – 517. IEEE Computer Society, November 2004.
- [SP97] L. Song and K. K. Parhi. Low-energy digit-serial/parallel finite field multipliers. *Journal of VLSI Signal Processing Systems*, 2(22):1–17, 1997.

Metrics of Password Management Policy

Carlos Villarrubia, Eduardo Fernández-Medina, and Mario Piattini

Alarcos Research Group,
Information Systems and Technologies Department,
UCLM-Soluziona Research and Development Institute,
University of Castilla-La Mancha,
Paseo de la Universidad, 4 – 13071 Ciudad Real, Spain
{Carlos.Villarrubia, Eduardo.FdezMedina, Mario.Piattini}@uclm.es

Abstract. The necessity to management the computer security of an institution implies an evaluation phase and the most common method to carry out this evaluation it consists on the use of a set of metrics. As any system of information needs of an authentication mechanism being the most used one those based on password, in this article we propose a set of metric of password management policies based on the most outstanding factors in this authentication mechanism. Together with the metrics, we propose a quality indicator derived from these metrics that allows us to have a global vision of the quality of the password management policy used and a complete example of calculation of the proposed metric. Finally, we will indicate the future works to be performed to check the validity and usefulness of the proposed metrics.

Keywords: Security management, assurance, metrics, passwords.

1 Introduction

Information and its support processes together with systems and nets are important resources for any organization. These resources are continuously subjected to risks and insecurities coming from a great variety of sources, where there are threats based on malicious code, programming errors, human errors, sabotages or fires.

This concern has encouraged many organizations and researchers to propose several metrics to evaluate security of their information systems. In general, there is a consensus regarding the fact that choosing these metrics depends on the concrete security need of each organization. The majority of performed proposals put forward methods to choose these metrics [1, 4, 19, 22, 26, 27]. In addition, sometimes, it is suggested the need of developing specific methodologies for each organization [7].

In any proposal, the need is to quantify the different security aspects to be able to understand, control, and improve confidence in the information system.

If an organization does not use security metrics for its decision making process, the choices will be motivated by subjective aspects, external pressures and even purely commercial motivations.

With the purpose of systematizing all these proposals, we have developed a classification outline of security metrics [29] where the proposed metrics in the existing literature have been included. In our work, we will conclude that the majority

of proposed metrics are general. This class of metrics only measure generic actions related to security and in an indirect way, specific objectives such as confidentiality, integrity and availability.

1.1 Authentication Systems

The use of an authentication system requires the integration of multiple elements; depending on the used techniques, it is necessary to use cryptography, medicine, psychology, systems analysis and protocol design. All authentication systems are designed to assure the identity of a participant to other participant and this process requires that the first participant demonstrates his identity according to any kind of information (knowledge evidence, possession evidence, and biological evidence). This authentication evidence can be a word or a password as it is used in the majority of operating systems and applications (knowledge evidence), a cryptographic card (possession evidence) or any biological characteristic of the individual to be authenticated and that is measured through a biometric device (biological evidence).

Historically, the use of a mechanism based on passwords has been the most used method. The importance of this authentication mechanism has led to the elaboration of rules and recommendations of multiple levels [11, 12, 13, 14, 20, 21]. The fact that this method is very easy to use in all systems together with its low cost has motivated this acceptance [18]. Deficiencies of this method have been widely studied and measures have been proposed to limit these disadvantages [2, 9, 23]. In some designs, the main disadvantages are linked to the necessary confidence in users when dealing with passwords while in other occasions, these disadvantages are motivated by designs that assumed a secure environment (such as, intranets) and that have been used in other environments (for example, the Internet) [10].

All these problems should indicate that passwords are a mechanism to be replaced but the users' acceptance of their use, their low cost together with the complexity and costs of the alternative methods guarantee their short and medium term continuance.

In this paper, we will propose metrics and indicators related to the password management policy due to the lack of specific proposals in special relevant areas in information system security.

In section 2, we will propose password management policy metrics justifying why they are necessary and classifying the proposed set according to several criteria. In section 3, we will put forward a classification according to levels of password management policies that allow organizations to know their current situation, to propose the relevant improvement and to relate comparisons between different institutions to know the best practices. Finally, we will present some of the obtained conclusions and a proposal of future work in this field.

2 Proposal of Password Management Metrics

The methodology used to derive the password management metrics has consisted on a study of all the factors that intervene in the password management. For this purpose, it has been gathered of the existent literature these factors [2,3,9,13,18,20,21]. These metrics do not try to cover the whole problem but to capture the most representative problems. In this hypothesis, it is not included the use of passwords for the authentication between processes or hosts. On the contrary, it is only studied the participation of a

person as an entity to be authenticated. Multifactor authentication systems where one of the authentication mechanisms is a password are not included either.

The definition of these metrics will be performed by defining the following aspects:

- *Name*: Representative name of the metric.
- *Description of the metric*: Generally, it describes the name of the metric by indicating the method to calculate values.
- *Life cycle phase*: For a better understandability and analysis of measures, metrics are classified according to their role within the life cycle of passwords.
 - *General*: Those metrics that could be in two or more phases are included.
 - *Assignment*: All metrics related to the assigning of initial identifiers and passwords to the users are included.
 - *Storage*: It contemplates the problem of storing passwords by the system.
 - *Transmission*: It includes the metrics related to the authentication protocols used by the user or the communication of the password to the user by the authentication system.
 - *Use*: Metrics that measure the way of use of the password by the user.
 - *Renewal*: Area of metrics related to the password modification.
- *Scale*: Set of values associated of this metric.
- *Multivalued*: Some of this metrics are susceptible of having several simultaneous measures. With this attribute it is indicated if the metric can have or not several simultaneous measures.

The names, description of the metrics, life cycle phase and multivalued that we have considered are as follows:

Table 1. Password management metrics

Name	Description	Phase	Mult.
Users Training	Type of training received by users for dealing with and selecting, if it is the case, passwords.	General	Yes
Group Password	Existence of passwords used by a group of users or passwords necessary to access to resources that do not have an access control mechanism separated from the authentication mechanism.	General	No
Action Register	Type of register used by the information system to monitor the actions related to the password management.	General	Yes
Alphabet Size	Number of characters of the alphabet used for the creation of passwords valid in the system.	Assignment	No
Number of Different Classes Demanded	Number of classes which the alphabet is divided into and that are required to determine a valid password.	Assignment	No
Minimum Length	Number of minimum characters required for a valid password.	Assignment	No
Source Selection	Set of agents that can be used to choose a password.	Assignment	No
Selection Restriction	Set of restrictions that avoid that the selection source uses a password easy to be found out by third parties.	Assignment	Yes
User Identifier Class	Type of user identifier used by the system.	Assignment	No
Predefined Users	Treatment that predefined users receive from the system.	Assignment	Yes

Table 1. (continued)

Storage Class	Way of passwords storage in the authentication system.	Storage	Yes
Initial Communication	Method of communication of the initial password or a re-assignment of the user by the authentication system.	Transmission	Yes
Net Transmission	Mechanism of transmission used by the authentication protocol for the confidentiality and integrity of password.	Transmission	No
Input Visualization	Method used by the system for the visualization of the password when it is required to the user.	Use	No
Maximum Number of Erroneous Attempts	Maximum number of failed attempts before the authentication system makes a defense operation because of the risk of identity usurpation by a third party.	Use	No
Information about Use	Group of mechanisms used by the authentication system to inform the user about the authentications performed in the past.	Use	No
Authentication Period	Maximum time after which the access control asks for a user re-authentication.	Use	No
Block by User Cancellation	Procedures used to guarantee that users that were legitimate in the past, are avoided to access the system.	Renewal	No
Minimum Life Time	Minimum life time of a valid password.	Renewal	No
Maximum Life Time	Maximum life time of a valid password. When this time goes by, the user is forced to change the password.	Renewal	No
Record Length	Number of valid passwords used by the user in the past and that the system does not allow to reuse.	Renewal	No
Password Reassigning	Procedure used to reactivate the credential of a user that does not remember his password.	Renewal	No

3 Indicator of Level of Security in the Password Management

The definition of a set of metrics is not enough for an organization to be able to use them to manage the necessary changes in the field of those metrics. It is necessary to have information about the way of use and the impact of the values of the metrics on the system management.

With this objective, we have proposed some pre-established values for each metric that facilitates its use. Except for some of them, these values are ordered according to a hierarchy, starting by a minimum value to a maximum one, passing through intermediate values in the majority of metrics. When an organization has a superior value in each metric, it will have a higher confidence in its authentication system.

As a general principle of computer security, it is not generally adequate to increase the values in some metrics without a generalized increase in all of them. Taking this principle as an objective, it is proposed an indicator of quality of password management policy based on five levels. This proposal is based on the usefulness shown in the maturity models and in the metrics management programmes [5, 6, 8, 26].

These levels are structured from a minimum level (level 1) to a maximum level (level 5). The values required in each metric are defined in each level. In some of these metrics, it is also defined a recommended value for each level. These

recommendations have the purpose of providing the indicator with flexibility, making it possible to define the required values at the lowest possible measure in each level.

When a metric has several values demanded in a level, this indicates that all those values should be had to consider that level has been reached. When in a metric it is demanded the same values in several levels, it is considered that the metric is in the higher level.

Anyway, the character of having recommended in a value of a metric does not have influence in its level and only has meaning for the calculation of the value of the indicator of quality of password management like it is described later on this section. Finally, the value '+' it indicates that the value of that metric in that level is overcome because this metric have a bigger value that the one demanded or recommended for that level. Table 2 shows our analysis for each metric, considering the above-mentioned levels.

Table 2. Values of metric and associate level

Users Training (Multivalued)	Level 1	Level 2	Level 3	Level 4	Level 5
No Training	Oblig. ¹				
Information when the user registration is made	Rec. ²	Oblig.	Oblig.	Oblig.	Oblig.
Compulsory course	Rec.	Rec.	Rec.	Oblig.	Oblig.
Periodic course	+ ³	+	+	Rec.	Oblig.
Group Password	Level 1	Level 2	Level 3	Level 4	Level 5
Existence of group passwords or access to resources passwords	Oblig.				
Unique existence of a group of administrators	+	Oblig.	Oblig.		
There are not group passwords	+	+	Rec.	Oblig.	Oblig.
Action Register (Multivalued)	Level 1	Level 2	Level 3	Level 4	Level 5
No action register	Oblig.				
Registration register	+	Oblig.	Oblig.	Oblig.	Oblig.
Renewal and cancellation register	+	Rec.	Oblig.	Oblig.	Oblig.
Block and re-assignment register	+	Rec.	Rec.	Oblig.	Oblig.
Alphabet Size	Level 1	Level 2	Level 3	Level 4	Level 5
Less than or equal to ten characters	Oblig.				
Between eleven and twenty-five characters	+	Oblig.			
Between twenty-six and fifty characters	+	Rec.	Oblig.		
Between fifty-one and seventy-five characters	+	+	Rec.	Oblig.	
More than seventy-five characters	+	+	Rec.	Rec.	Oblig.
Number of Different Classes demanded	Level 1	Level 2	Level 3	Level 4	Level 5
One	Oblig.				
Two	+	Oblig.			
Three	+	+	Oblig.	Oblig.	
Four or more	+	+	+	Rec.	Oblig.
Minimum Length	Level 1	Level 2	Level 3	Level 4	Level 5
Less than or equal to four characters	Oblig.				
Between five and eight characters	+	Oblig.			
Between nine and twelve characters	+	+	Oblig.		
Between thirteen and sixteen characters	+	+	+	Oblig.	
Greater than sixteen characters	+	+	+	+	Oblig.

¹ Oblig. Obligatory value.

² Rec.: Recommended value.

³ +: Overcome value in this level.

Table 2. (continued)

Source Selection	Level 1	Level 2	Level 3	Level 4	Level 5
User	Oblig.	Oblig.	Oblig.	Oblig.	Oblig.
System	+	+	+	Rec.	Rec.
Selection Restriction (Multivalue)	Level 1	Level 2	Level 3	Level 4	Level 5
No restriction					
User information	Oblig.	Oblig.	Oblig.	Oblig.	Oblig.
Keys combinations	+	Rec.	Rec.	Oblig.	Oblig.
Dictionary password	+	+	Rec.	Oblig.	Oblig.
Variations of the previous ones	+	+	+	Rec.	Oblig.
User Identifier Class	Level 1	Level 2	Level 3	Level 4	Level 5
Public identifier	Oblig.	Oblig.	Oblig.		
Semi-public identifier	+	+	Rec.	Oblig.	
Private identifier	+	+	+	Rec.	Oblig.
Predefined Users (Multivalue)	Level 1	Level 2	Level 3	Level 4	Level 5
No change					
Password change	Oblig.	Oblig.	Oblig.	Oblig.	Oblig.
Identifier change	+	+	Rec.	Oblig.	Oblig.
Storage Class (Multivalue)	Level 1	Level 2	Level 3	Level 4	Level 5
Clear storage					
Irreversible storage	Oblig.	Oblig.	Rec.	Rec.	Rec.
Encrypted storage	+	+	Oblig.	Oblig.	Oblig.
Initial Communication (Multivalue)	Level 1	Level 2	Level 3	Level 4	Level 5
Non-secure transmission	Oblig.				
Transmission with compulsory change of password	+	Oblig.	Oblig.	Rec.	Rec.
Secure transmission	+	+	Rec.	Oblig.	Oblig.
Net Transmission	Level 1	Level 2	Level 3	Level 4	Level 5
Clear transmission					
Use of a challenge-response protocol	Oblig.	Oblig.	Oblig.		
Encrypted transmission	+	+	Rec.	Oblig.	Oblig.
Input Visualization	Level 1	Level 2	Level 3	Level 4	Level 5
Clear visualization					
Visualization of number of characters	Oblig.	Oblig.	Oblig.		
No visualization	+	+	Rec.	Oblig.	Oblig.
Maximum Number of Erroneous Authentication Attempts	Level 1	Level 2	Level 3	Level 4	Level 5
No limit	Oblig.				
Between eleven and fifty attempts	Rec.	Oblig.			
Between four and ten attempts	+	Rec.	Oblig.	Oblig.	
Less than or equal to three attempts	+	+	+	Rec.	Oblig.
Information about Use	Level 1	Level 2	Level 3	Level 4	Level 5
No information	Oblig.	Oblig.	Oblig.	Oblig.	Oblig.
Information about the last use	+	+	Rec.	Rec.	Rec.
Authentication Period	Level 1	Level 2	Level 3	Level 4	Level 5
Work session	Oblig.	Oblig.			
Maximum of fifteen minutes inactivity	+	+	Oblig.	Oblig.	
Maximum of five minutes inactivity	+	+	+	Rec.	Oblig.
Block by User Cancellation	Level 1	Level 2	Level 3	Level 4	Level 5
Without an established method	Oblig.				
Periodic elimination (maximum of six months period)	Rec.	Oblig.	Oblig.		
Time limit established during registration	+	+	Rec.	Oblig.	Oblig.
Minimum Life Time	Level 1	Level 2	Level 3	Level 4	Level 5
There is not minimum life time	Oblig.	Oblig.	Oblig.	Oblig.	
There is a minimum life time (equal to or greater than 1 day)	+	+	Rec.	Rec.	Oblig.

Table 2. (continued)

Maximum Life Time	Level 1	Level 2	Level 3	Level 4	Level 5
Greater than twelve months	Oblig.				
Lower than or equal to twelve months	+	Oblig.			
Lower than or equal to six months	+	+	Oblig.	Oblig.	
Lower than or equal to three months	+	+	+	Rec.	Oblig.
Record Length	Level 1	Level 2	Level 3	Level 4	Level 5
One	Oblig.				
Lower than or equal to three	+	Oblig.			
Lower than or equal to ten	+	+	Oblig.		
Lower than or equal to twenty-five	+	+	+	Oblig.	
Greater than twenty-five	+	+	+	+	Oblig.
Password Reassigning	Level 1	Level 2	Level 3	Level 4	Level 5
The previous password is reassigned	Oblig.				
A new password is assigned	Rec.	Oblig.	Oblig.	Oblig.	Oblig.

The calculation of the value of the indicator of quality of the password management policy requires them to be had as minimum the values of the metric ones with the requirement of obligatory, overcome or recommended. It is necessary to highlight that although the number of metric is twenty-two, the obtained values can be greater because several metric they can have several values simultaneously (for example, users training). The minimum number of values to reach the corresponding level is shown in the table 3.

Table 3. Number of values in each level

Level	Minimum number
1	22
2	22
3	23
4	28
5	30

3.1 Application of Metrics

In this section a concrete case of application of metric is detailed. The used system of information has the following characteristic: the new user is informed the password management policy and in the maximum term of one month he receives a formation session where aspects of computer security are included. The election of password carries out it the user with the following restrictions: 8 minimum characters of an alphabet with discrimination between uppercase and lowercase and with a mixture of digits. In the communication of the initial password to the user puts under an obligation to this to a change of password and these they are stored encrypted and using a dispersion function to be irreversible. These characteristics together with others that are deduced from the table 4 allow us to obtain the following values for the metric.

Table 4. Values of each metric in the example

Metric: Value	Level 1	Level 2	Level 3	Level 4	Level 5
Users Training: Information when the user registration is made	Rec.	Oblig.	Oblig.	Oblig.	Oblig.
Users Training: Compulsory course	Rec.	Rec.	Rec.	Oblig.	Oblig.
Group Password: Unique existence of a group of administrators	+	Oblig.	Oblig.		
Action Register : Registration register	+	Oblig.	Oblig.	Oblig.	Oblig.
Alphabet Size : More than seventy-five characters	+	+	Rec.	Rec.	Oblig.
Number of Different Classes demanded: Two	+	Oblig.			
Minimum Length: Between five and eight characters	+	Oblig.			
Source Selection: User	Oblig.	Oblig.	Oblig.	Oblig.	Oblig.
Selection Restriction: User information	Oblig.	Oblig.	Oblig.	Oblig.	Oblig.
User Identifier Class: Public identifier	Oblig.	Oblig.	Oblig.		
Predefined Users: Password change	Oblig.	Oblig.	Oblig.	Oblig.	Oblig.
Storage Class: Irreversible storage	Oblig.	Oblig.	Rec.	Rec.	Rec.
Storage Class: Encrypted storage	+	+	Oblig.	Oblig.	Oblig.
Initial Communication: Transmission with compulsory change of password	+	Oblig.	Oblig.	Rec.	Rec.
Net Transmission: Encrypted transmission	+	+	Rec.	Oblig.	Oblig.
Input Visualization: Visualization of number of characters	Oblig.	Oblig.	Oblig.		
Maximum Number of Erroneous Authentication Attempts: Between eleven and fifty attempts	Rec.	Oblig.			
Information about Use: No information	Oblig.	Oblig.	Oblig.	Oblig.	Oblig.
Authentication Period: Work session	Oblig.	Oblig.			
Block by User Cancellation: Periodic elimination (maximum of six months period)	Rec.	Oblig.	Oblig.		
Minimum Life Time: There is not minimum life time	Oblig.	Oblig.	Oblig.	Oblig.	
Maximum Life Time: Lower than or equal to twelve months	+	Oblig.			
Record Length: Lower than or equal to ten	+	+	Oblig.		
Password Reassigning: A new password is assigned	Rec.	Oblig.	Oblig.	Oblig.	Oblig.

With these measures the table 5 is obtained with a summary for level and for the obligatory, recommended or overcome character of each metric.

The table 4 shows that the used password management policy has the levels 1 and 2 because has all the required values. However, to obtain the level 3 he has to improve in four metrics: number of different classes demanded, minimum length,

Table 5. Values for level in the example

Total	Level 1	Level 2	Level 3	Level 4	Level 5
Obligatory value	9	18	16	13	12
Recommended value	4	2	4	3	2
Overcome value	11	4			
<i>Total of values</i>	24	24	20	16	13

authentication period and maximum life time. Finally, to reach the level 4 he needs to improve in eight metrics and for the level 5 in ten metrics.

4 Conclusions and Future Work

In this work, we have proposed a set of metric and an indicator of level of security in password management policy that they complete the objective of evaluating the authentication process through passwords.

We have proposed twenty-two metrics grouped into six areas covering the whole cycle of password management. Due to the diversity of these metrics, where some of them have a potentially infinite value range (for instance, password length) and others have a very limited value range (for instance, password reassignment), the definition of metrics includes a limited set of values that simplifies the process of obtaining measures and the use of metrics for decision making.

As a method of global valuation of the password management policy, it is proposed an indicator of quality whose range of values is formed by five levels. This indicator makes it possible to inform, in a single and comprehensible way, all actors involved in the organization security about the level of quality reached in an information system.

It is included one application example, a supposition where the level of each metric one is obtained together with the indicator of level of security of the whole group of metric. In this supposition we show the simplicity in the orientation to the manager to direct their future actions.

This proposal is made within the framework of a wider project of metrics definition that studies all security general areas. Nevertheless, in the area of identification and authentication, it is necessary to extend these metrics to the exploitation of the information system to complete the password management system.

Furthermore, the majority of organizations have a diversity of information systems with different requirements as well as different authentication mechanisms. To obtain an overall vision, through a set of metrics, it is necessary to combine all this information in a coherent and useful way for the organization board of directors and technical staff. In this aspect, the proposed metrics must be completed with others taking into account these circumstances.

Finally, we intend to be carried out like future works a study of the password management policy of a group of organizations selected to check the utility of the metric proposals, to validate the proposed group and to be a reference in best practices in this environment.

Acknowledgements

This research is part of the DIMENSIONS projects, partially financed by the FEDER and the Consejería de Educación y Ciencia de la Junta de Comunidades de Castilla-La Mancha (PBC-05-012-1), CALIPO (TIC2003-07804-C05-03) and RETISTIC (TIC2002-12487-E) granted by the “Dirección General de Investigación del Ministerio de Ciencia y Tecnología” (Spain).

References

1. ACSA, editor. *Proceedings of the Workshop on Information Security System Scoring and Ranking*, Williamsburg, Virginia, may 2001.
2. A. Adams, M. A. Sasse, and P. Lunt. *Making passwords secure and usable*. In *Proceedings of Human Computer Interaction*, Bristol, England, aug 1997.
3. M. Bishop. *Comparing authentication techniques*. In *Proceedings of the Third Workshop on Computer Incident Handling*, pp. 1–10, aug 1991.
4. P. Bouvier and R. Longeon. *Le tableau de bord de la sécurité du système d'information*. *Sécurité Informatique*, jun 2003.
5. Carnegie Mellon University, Pittsburgh, Pennsylvania. *SSE-CMM Model Description Document, 3.0 edition*, jun 2003.
6. D. A. Chapin and S. Akridge. *How can security be measured?* *Information Systems Control Journal*, 2:43–47, 2005.
7. C. Colado and A. Franco. *Métricas de seguridad: una visión actualizada*. *SIC. Seguridad en Informática y Comunicaciones*, 57:64–66, nov 2003.
8. Department of the Air Force. *AFI33-205. Information Protection Metrics and Measurements Program*, aug 1997.
9. A. Halderman, B. Waters, and E. W. Felten. *A convenient method for securely managing passwords*. In *Proceedings of the 14th International World Wide Web Conference*, pp. 471–479, Chiba, Japan, may 2005.
10. ISO. *ISO 7498-2. Open Systems Interconnection - Basic Reference Model - Part 2: Security Architecture*, 1989.
11. ISO/IEC. *ISO/IEC TR 13335-1. Guidelines for the Management of IT Security. Part I: Concepts and Models of IT Security*, 1996.
12. ISO/IEC. *ISO/IEC 15408. Evaluation Criteria for IT Security*, dec 1999.
13. ISO/IEC. *ISO/IEC 17799. Code of Practice for Information Security Management*, 2000.
14. G. King. *Best security practices: An overview*. In *Proceedings of the 23rd National Information Systems Security Conference*, Baltimore, Maryland, oct 2000. NIST.
15. J. M. Marcelo. *Seguridad de las Tecnologías de la Información, capítulo Identificación y Evaluación de Entidades en un Método AGR*, pp. 69–103. AENOR, 2003.
16. W. L. McKnight. *What is information assurance?* *CrossTalk. The Journal of Defense Software Engineering*, pp. 4–6, jul 2002.
17. R. T. Mercuri. *Analyzing security costs*. *CACM*, 46(6):15–18, jun 2003.
18. R. Morris and K. Thompson. *Password security: A case history*. *CACM*, 22(11):594–597, 1979.
19. F.Nielsen. *Approaches of security metrics*. Technical report, NIST-CSSPAB, jun 2000.
20. NIST. *FIPS-112: Password Usage*, may 1985.
21. NIST. *FIPS-181: Automated Password Generator*, oct 1993.
22. S. C. Payne. *A guide to security metrics*. Technical report, SANS Institute, jul 2001.
23. B. Pinkas and T. Sander. *Securing passwords against dictionary attacks*. In *Proceedings of the ACM Computer and Security Conference (CSC' 02)*, pp. 161–170, nov 2002.
24. G. Schuedel and B. Wood. *Adversary work factor as a metric for information assurance*. In *Proceedings of the New Security Paradigm Workshop*, pp. 23–30, Ireland, sep 2000.
25. M. Swanson. *Security self-assessment guide for information technology systems*. Tech. Report NIST 800-26, National Institute of Standards and Technology, nov 2001.
26. M. Swanson, N. Bartol, J. Sabato, . J. Hash, and L. Graffo. *Security metrics guide for information technology systems. Technical Report NIST 800-55*, National Institute of Standards and Technology, jul 2003.

27. R. B. Vaughn, Jr., R. Henning, and A. Siraj. *Information assurance measures and metrics – state of practice and proposed taxonomy*. In Proceedings of the 36th Hawaii International Conference on Systems Sciences, 2003.
28. R. B. Vaughn, Jr., A. Siraj, and D. A. Dampier. *Information security system rating and ranking*. CrossTalk. The Journal of Defense Software Engineering, pp. 30–32, may 2002.
29. C. Villarrubia, E. Fernández-Medina, and M. Piattini. *Towards a classification of security metrics*. In Proceedings of the 2nd international workshop on security in information systems (WOSIS 2004), pp. 342–350, apr 2004.

Using UML Packages for Designing Secure Data Warehouses

Rodolfo Villarroel¹, Emilio Soler², Eduardo Fernández-Medina³, Juan Trujillo⁴,
and Mario Piattini³

¹ Departamento de Computación e Informática. Catholic University of Maule,
Avenida San Miguel 3605 Talca, Chile
rvillarr@spock.ucm.cl

² Departamento de Informática. University of Matanzas,
Autopista de Varadero Km. 3. Matanzas, Cuba
emilio.soler@umcc.cu

³ Alarcos Research Group. Information Systems and Technologies Department,
UCLM-Soluziona Research and Development Institute,
University of Castilla-La Mancha
Paseo de la Universidad, 4 - 13071 Ciudad Real, Spain
{Eduardo.FdezMedina, Mario.Piattini}@uclm.es

⁴ Departamento de Lenguajes y Sistemas Informáticos. University of Alicante,
C/San Vicente S/N 03690 Alicante, Spain
jtrujillo@dlsi.ua.es

Abstract. Due to the sensitive data contained in Data Warehouses (DWs), it is essential to specify security measures from the early stages of the DWs design and enforce them. In this paper, we will present a UML profile to represent multidimensional and security aspects of our conceptual modeling. Our approach proposes the use of UML packages in order to group classes together into higher level units creating different levels of abstraction, and therefore, simplifying the final model. Furthermore, we present an extension of the relational model to consider security and audit measures represented in the conceptual modeling. To accomplish this, we based on the Relational Package of the Common Warehouse Metamodel (CWM) and extend it to properly represent all security and audit rules defined in the conceptual modeling of DWs. Finally, we will show an example to illustrate the applicability of our proposal.

1 Introduction

Organizations depend increasingly on information systems, which rely upon databases and data warehouses (DWs), which need increasingly more quality and security. Indeed, the very survival of organizations depends on the correct management, security and confidentiality of information [2]. In fact, as some authors have remarked [1, 4], information security is a serious requirement which must be carefully considered, not as an isolated aspect, but as an element which turns up as an issue in all stages of the development lifecycle, from the requirement analysis to implementation and maintenance. As other authors point out [6, 9], even though most

DWs are implemented into relational DBMS, security measures and access control models specified for transactional (relational) databases are not appropriate for DWs. The main reason is that the security measures for DWs must be defined on a multidimensional basis, since DW users query the DW in terms of facts, dimensions, classification hierarchy levels and so on.

In MD modeling, information is structured into facts and dimensions. A fact represents interesting measures of a business process (sales, deliveries, etc.), whereas a dimension considers the context for analyzing a fact (product, customer, time, etc.). A high number of dimensions with their corresponding hierarchies, and a considerable number of facts sharing dimensions and classification hierarchies will lead to a very complex design, thereby increasing the difficulty in reading the modeled system. Therefore, a secure MD conceptual model should also provide techniques to avoid flat diagrams to simplify the final model.

In this paper, we present a UML profile to represent MD and security aspects of our conceptual modeling. We propose the use of UML packages in order to group classes together into higher level units creating different levels of abstraction. Furthermore, we present an extension of the relational model, aligned with OMG, to consider security and audit measures represented in the conceptual modeling.

The remainder of this paper is structured as follows. In Section 2 we summarize the main related work. In Section 3 we present the Common Warehouse Metamodel (CWM) and the four-layer architecture of OMG. In Section 4 we present the UML 2.0 profile for secure multidimensional modeling. In Section 5 we present an extension of the relational metamodel of CWM. In Section 6 we state an example to support the conceptual design of secure data warehouses using packages. Finally, in Section 7, we draw some conclusions and sketch our immediate future work.

2 Related Work

In the past few years, several approaches have been proposed for representing the main multidimensional (MD) properties at the conceptual level [5, 11-13]. Nevertheless, none of these approaches for MD modeling, considers security to be an important issue in their conceptual models, so they do not solve the problems arising from this question in these kinds of systems. It is true that, in the relevant literature, we can find several initiatives for the inclusion of security in data warehouses [6, 9, 10]. However, none of them considers security aspects which incorporate all stages of the system development cycle, nor the introduction of security into MD design.

The previous work presented in [8] introduced a Model Driven Architecture (MDA) oriented framework for the DW development, choosing the ROLAP (Relational On-Line Analytical Processing) like DBMS and the Platform Specific Model (PSM) is modeled by using the relational metamodel from the CWM. However, none security and audit measures can be modeled in this metamodel.

To the best of our knowledge, only our previous works [3, 14] sets the basis for providing a conceptual model for the design of secure DWs. In this paper, our previous works are refined, adapting our UML profile for a secure multidimensional modeling to the proposal for the MD modeling with UML package diagrams [7].

3 CWM and the Four Layer Architecture of OMG

The standard OMG (Object Management Group) promotes the theory and practice of object-oriented technology in software development, based on the four-layer metamodel architecture. A model at one layer is used to specify models in the layer above. The four-layer architecture is shown in Table 1.

Table 1. The four-layer architecture of OMG

Meta-level	MOF Terms	Examples
M3	Meta-metamodel	The MOF model
M2	Metamodel, metadata	UML metamodel, CWM metamodel
M1	Model, metadata	UML models, CWM metadata
M0	Object, data	Modeled systems, data warehouse

The main purpose of the CWM is to enable easy interchange of warehouse and business intelligence metadata between warehouse tools, warehouse platforms and warehouse metadata repositories in distributed heterogeneous environments.

CWM is organized in 21 separate packages which they were grouped into five stackable layers by means of similar roles (see Fig. 1).

Management	Warehouse Process			Warehouse Operation		
Analysis	Transformation	OLAP	Data Mining	Information Visualization	Business Nomenclature	
Resource	Object	Relational	Record	Multidimensional		XML
Foundation	Business Information	Data Types	Expressions	Keys and Indexes	Software Deployment	Type Mapping
Object Model	Core		Behavioral	Relationships		Instance

Fig. 1. CWM metamodel layering and its packages

From the organization represented in Fig. 1, we will mainly focus our work (for a secure relational modeling of DWs) on the Resource layer and, more precisely, on the Relational package as a relational metamodel to describe that represent metadata of relational data resources.

4 A UML profile for Secure Multidimensional Modeling

The goal of this UML profile is to be able to design a MD conceptual model, but classifying information at the same time, in order to define which properties the user has to have in order to be entitled to gain access to information. We can define, for each element of the model (fact class, dimension class, fact attribute, etc.), its security information, specifying a sequence of security levels, a set of user compartments and a set of user roles. We can also specify security constraints considering these security attributes. Our profile will be called SECDW (Secure Data Warehouses) and will be represented as a UML package. This profile will not only inherit all properties from

the UML metamodel but it will also incorporate new data types, stereotypes, tagged values and constraints. In Fig. 2, a high-level view of our SECDW profile is provided.

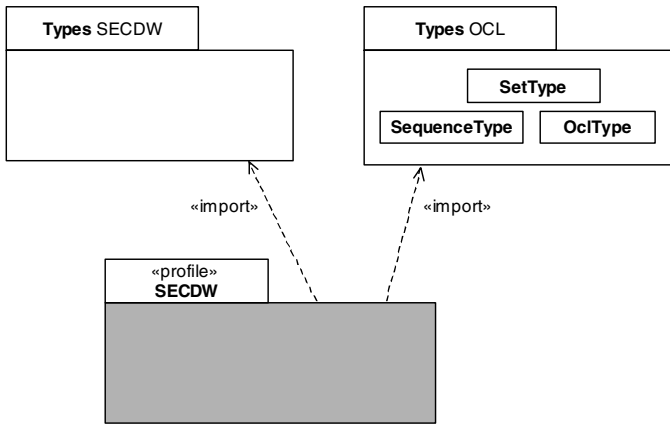


Fig. 2. High level view of our SECDW profile

We have defined a package that includes all the stereotypes that will be necessary in our profile (see Fig. 3).

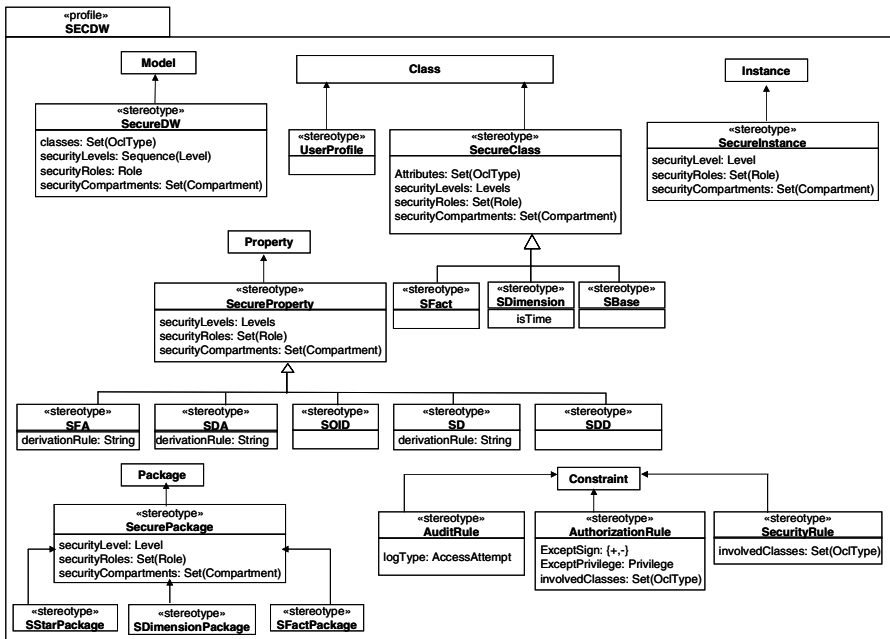


Fig. 3: New stereotypes

This profile contains four types of stereotypes:

- Secure Class, secure package and secure data warehouses stereotypes (and stereotypes inheriting information from them) that contain tagged values associated with attributes (model or class attributes), security levels, user roles and organizational compartments.
- Attribute stereotypes (and stereotypes inheriting information from attributes) and instances, which have tagged values associated with security levels, user roles and organizational compartments.
- Stereotypes that allow us to represent security constraints, authorization rules and audit rules.
- UserProfile stereotype, which is necessary to specify constraints depending on particular information of a user or a group of users.

4.1 Using UML Packages for Secure Multidimensional Modeling

In our approach, the main structural properties of MD models are specified by means of a UML class diagram in which the information is clearly separated into facts and dimensions. Our approach proposes the use of UML packages in order to group classes together into higher level units creating different levels of abstraction, and therefore, simplifying the final model.

The different levels show how one package can be further exploded by defining its corresponding elements into the next level as we describe as follows (see Fig. 4):

- **Level 1:** Model definition. A package represents a star schema of a conceptual MD model.
- **Level 2:** Star schema definition. A package represents a fact or a dimension of a star schema.
- **Level 3:** Dimension/fact definition. A package is exploded into a set of classes that represent the hierarchy levels defined in a dimension package, or the whole star schema in the case of the fact package.

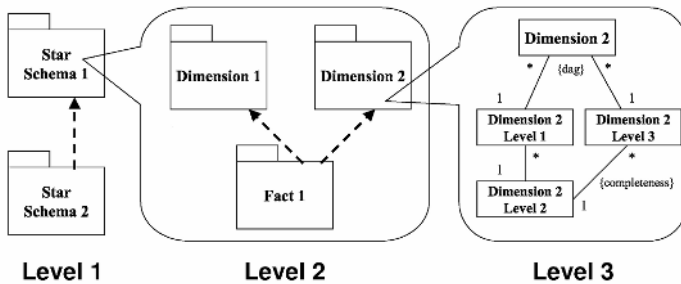


Fig. 4. Levels of a MD model explosion using packages

Next, we will present the metamodel of our OO conceptual MD approach using a UML class diagram. In order to simplify this diagram, we have divided it into three levels. In Fig. 5, the content of metamodel level1 package is shown. This

package specifies the modeling elements that can be applied at the metamodel level 1 of our approach. At this level, only the StarPackage model element is allowed.

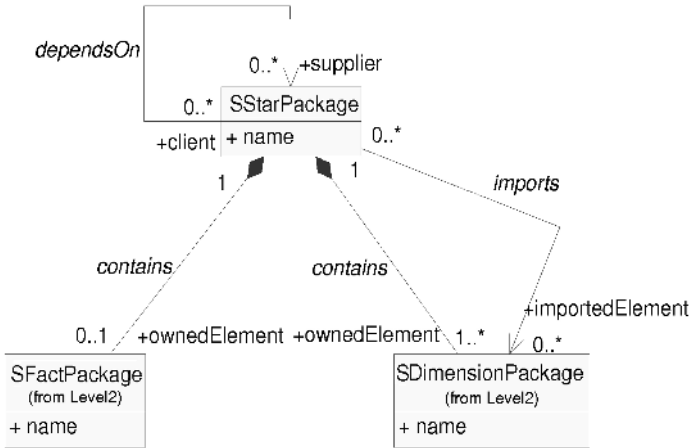


Fig. 5. Metamodel: level 1

In Fig. 6, we will show the content of metamodel level2 package.

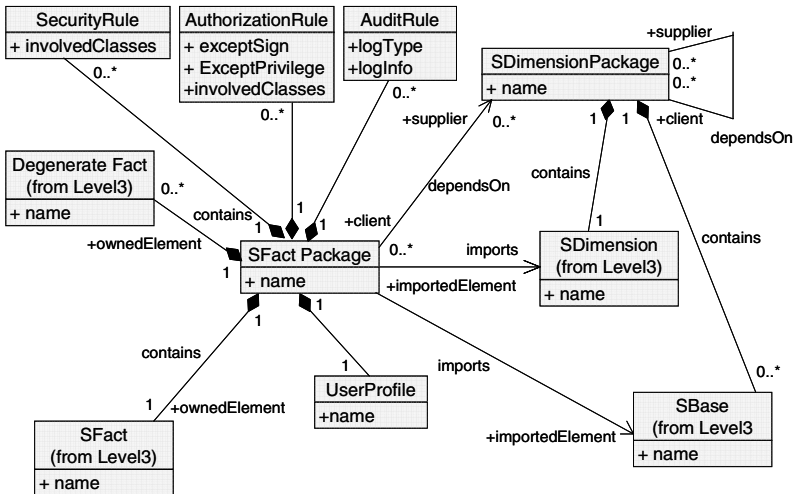


Fig. 6. Metamodel: level 2

The stereotypes SecurityRule, AuthorizationRule, and AuditRule can have the following information:

- SecurityRule (sensitivity information associated).
- AuthorizationRule (information to permit or deny access).
- AuditRule (information to analyze the user behaviour when using the system).

Finally, in Fig. 7, we will show the content of the metamodel level3 package. This diagram represents the main MD properties of our modeling approach.

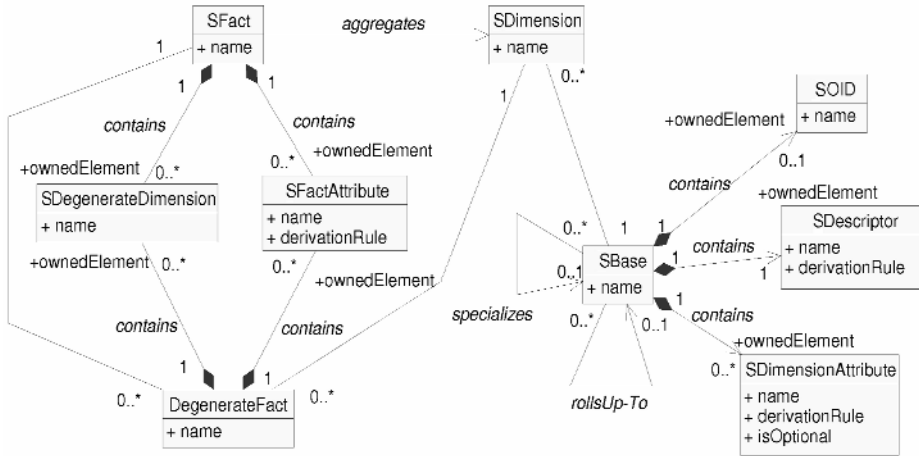


Fig. 7. Metamodel: level 3

5 Secure Multidimensional Modeling at the Logical Level

In this section we outline the relational metamodel of CWM. We only use part of the relational CWM metamodel for our purposes; which allow us to represent tables, columns, primary keys and foreign keys. However, for representing security and audit measures in the metamodel, we need to add some metaclasses. In Fig. 8 we show part of the relational CWM metamodel extended.

The Schema metaclass aim the security at the model level. SecurityProperty metaclass inherit from the Constraint metaclass and specializes as SecurityLevels, SecurityCompartments and SecurityRoles metaclasses. Furthermore, for representing security constraints, authorization rules and audit rules in the metamodel we add AUDconstraint class, ARconstraint class and AURconstraint class, which inherit from SecurityConstraint. For specify constraints depending on particular information of a user or a group of users, we introduce the userProfile metaclass. Finally, we need add associations of Table and Column metaclasses with the metaclasses introduced in order to establish security in attributes and tables. For express the constraints (AuditRule, AuthorizationRule and SecurityRule) modeled in SECDW metamodel using notes, we need to add a new attribute OCLConstraint in the SecurityConstraint metaclass.

corresponding to each one of the secure packages that will be later represented at the following level through dimensions and hierarchy levels.

In Fig. 10, we present a detailed vision of SFactPackage Admission. In this case, as it corresponds to a secure fact package, it is shown the star schema complete. If we had chosen a dimension, it will be only detailed the modeling (including security aspects) of the dimension with its hierarchy levels. In this figure, we can see the stereotypes for the Admission Fact Class, Diagnosis and Patient Dimensions and the classification hierarchies (or Base Class) corresponding to each dimension. The tagged values are represented as static security constraints at the class or attribute level. For example, we can see that the Admission Sfact class has the security levels from Secret to TopSecret and the user roles Health and Admin. At the attribute level, there is a cost static security rule that indicates that it can only be accessed by users having the admin. role. In addition, a series of UML notes can be seen, where dynamic security constraints (that depend on a condition), authorization rules for exceptions and audit rules are represented.

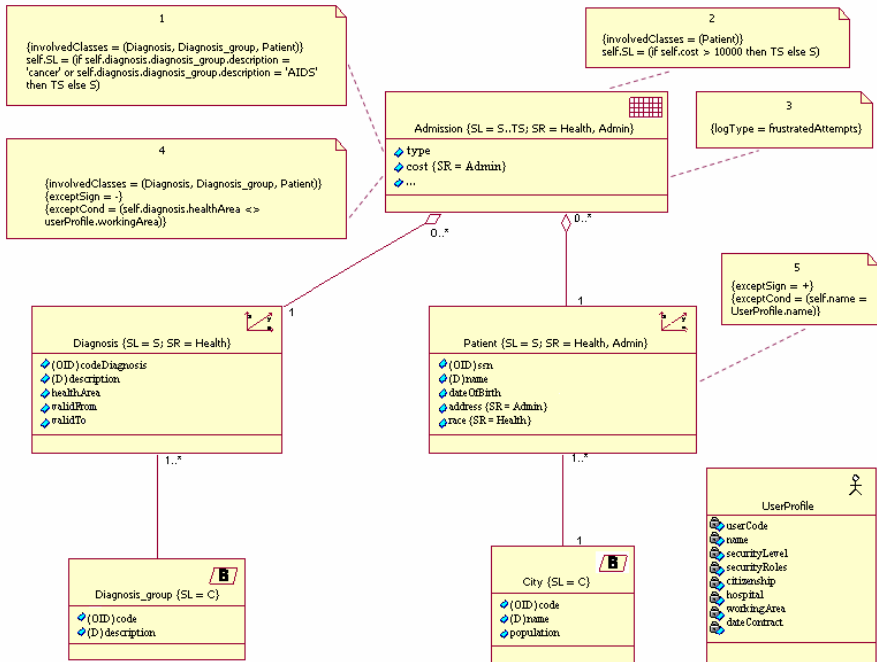


Fig. 10. Level 3: Content of SFactPackage Admission

7 Conclusions and Future Work

In this paper, we have presented a UML 2.0 profile for secure multidimensional modeling that extends previous works. To do so, we have used UML packages to represent our stereotypes, tagged values, and OCL constraints in our modeling.

Furthermore, we have presented an extension of the relational metamodel of the CWM in order to represent security and audit measures in the logical modeling of data warehouses.

Our immediate future work consists on the formal specification of all the required transformations between the conceptual and the logical models by using the Query-View-Transformation (QVT), thereby aligning our approach with the Model Driven Architecture. In this way, we will be able to specify all transformations in a formal language, thereby avoiding an arbitrary definition of these rules.

Acknowledgements

This research is part of the RETISTIC (TIC2002-12487-E) and METASIGN (TIN2004-00799) projects from the Spanish Ministry of Education and Science, the MESSSENGER (PCC-03-003-1) and DIMENSIONS (PBC-05-012-2) projects from the Regional Science and Technology Ministry of Castilla-La Mancha, the DADASMECA project (GV05/220) from the Regional Government of Valencia, and the COMPETISOFT project (506PI0297) financed by CYTED.

References

1. Devanbu, P. and Stubblebine, S.: *Software engineering for security: a roadmap*. in *Proceedings of the Conference on The Future of Software Engineering*. Ireland (2000)
2. Dhillon, G. and Backhouse, J.: *Information system security management in the new millennium*. Communications of the ACM (2000) **43**(7):125-128
3. Fernández-Medina, E., Trujillo, J., Villarroel, R., and Piattini, M.: *Extending the UML for Designing Secure Data Warehouses*. in *Int. Conference on Conceptual Modeling (ER 2004)*. Shanghai, China: Springer-Verlag. LNCS 3288 (2004)
4. Ferrari, E. and Thuraisingham, B.: *Secure Database Systems*, in *Advanced Databases: Technology Design*, Piattini, M. and Díaz, O., Editors, Artech House: London (2000)
5. Golfarelli, M., Maio, D., and Rizzi, S., *The Dimensional Fact Model: A Conceptual Model for Data Warehouses*. Int. Journal of Cooperative Information Systems (IJCIS), (1998) **7**.(2-3): 215-247.
6. Katic, N., Quirchmayr, G., Schiefer, J., Stolba, M., and Min Tjoa, A.: *A Prototype Model for Data Warehouse Security Based on Metadata*. in *9th Int. Workshop on Database and Expert Systems Applications (DEXA'98)*. Vienna. IEEE Computer Society (1998)
7. Luján-Mora, S., Trujillo, J., and Song, I.Y.: *Multidimensional Modeling with UML Package Diagrams*. in *Int. Conference on Conceptual Modeling - ER 2002*. Tampere, Finland: Springer. LNCS 2503 (2002)
8. Mazon, J., Trujillo, J., Serrano, M., and Piattini, M. Applying MDA to the development of data warehouses. in *8th ACM Int. Workshop on Data Warehousing and OLAP (DOLAP'05)*. Bremen, Germany (2005) 57-66
9. Priebe, T. and Pernul, G.: *Towards OLAP Security Design - Survey and Research Issues*. in *3rd ACM Int. Workshop on Data Warehousing and OLAP (DOLAP'00)*. USA (2000)
10. Rosenthal, A. and Sciore, E.: *View Security as the Basic for Data Warehouse Security*. in *2nd Int. Workshop on Design and Management of Data Warehouses (DMDW'00)*. Sweden (2000)

11. Sapia, C., Blaschka, M., Höfling, G., and Dinter, B.: *Extending the E/R Model for the Multidimensional Paradigm*. in *1st Int. Workshop on Data Warehouse and Data Mining (DWDM'98)*. Singapore: Springer-Verlag LNCS 1552 (1998)
12. Trujillo, J., Palomar, M., Gómez, J., and Song, I.Y., *Designing Data Warehouses with OO Conceptual Models*. IEEE Computer, special issue on Data Warehouses, 2001(34): 66-75.
13. Tryfona, N., Busborg, F., and Christiansen, J.: *starER: A Conceptual Model for Data Warehouse Design*. in *ACM 2nd Int. Workshop on Data Warehousing and OLAP (DOLAP'99)*. Missouri, USA: ACM (1999)
14. Villarroel, R., Fernandez-Medina, E., Trujillo, J., and Piattini, M.: *Towards a UML 2.0/OCL extension for Designing Secure Data Warehouses*. in *3rd. Int. Workshop on Security in Information Systems (WOSIS 2005)*. Miami, USA: INSTICC Press (2005)

Practical Attack on the Shrinking Generator^{*}

Pino Caballero-Gil¹ and Amparo Fúster-Sabater²

¹ D.E.I.O.C. University of La Laguna. 38271 La Laguna, Tenerife, Spain
pcaballe@ull.es

² Institute of Applied Physics. C.S.I.C. Serrano 144, 28006 Madrid, Spain
amparo@iec.csic.es

Abstract. This work proposes an efficient attack on the Shrinking Generator based on its characterization by means of Linear Hybrid Cellular Automata. The algorithm uses the computation of the characteristic polynomials of specific sub-automata and the generation of the Galois field associated to one of the Linear Feedback Shift Registers components of the generator. Both theoretical and empirical evidences for the effectiveness of the attack are given. The two main advantages of the described cryptanalysis are the determinism of bits prediction and the possible application of the obtained results to different generators.

1 Introduction

A binary additive stream cipher is a synchronous cipher in which the binary output of a keystream generator is added bitwise to the binary plaintext sequence producing the binary ciphertext. The main goal in stream cipher design is to produce random-looking sequences that are unpredictable in an efficient way. From a cryptanalysis point of view, a good stream cipher should be resistant against known-plaintext attacks.

Most known keystream generators are based on Linear Feedback Shift Registers (LFSRs). In particular, the Shrinking Generator (SG) is a nonlinear combinator based on two LFSRs such that the bits of one output are used to determine whether the corresponding bits of the second output are part of the overall keystream. Although there have been several approaches for attacking the SG, it produces pseudorandom sequences with good security properties. A basic divide-and-conquer attack requiring an exhaustive search through all the possible initial states and feedback polynomials of the selector LFSR was proposed in [14]. The authors of [7] described a correlation attack targeting the second LFSR. Another correlation attack based on searching specific subsequences of the output sequence was introduced in [8]. More recently, a distinguishing attack applicable when the second LFSR has a low-weight feedback polynomial was investigated in [5].

^{*} Research supported by the Spanish Ministry of Education and Science and the European FEDER Fund under Projects SEG2004-04352-C04-03 and SEG2004-02418.

On the other hand, Cellular Automata (CA) are discrete mathematical models in which a lattice of finite state machines, called cells, updates itself synchronously according to local rules [16]. CA have been proposed both for secret and public key cryptography [16], [12], [13]. Also cryptanalysis of certain CA based keystream generators have been published in [10] and [11]. Finally, in [6] a Cellular Automata-Based model for the Shrinking Generator was proposed. Such a work may be considered the starting point of this research.

This work has been laid out as follows. The next section gives relevant background about the basic structures we are dealing with: Linear Hybrid Cellular Automata and Shrinking Generators. The CA-Based model for the SG that is used in this work is described next. Section 4 gives the theoretical basis of the proposed CA-based cryptanalysis of the SG. Sections 5 and 6 introduce the full description of the algorithm and its analysis, respectively. Finally, in Section 7 several conclusions and open questions are drawn.

2 Preliminaries

Cellular automata are finite state machines that consist of arrays of n cells [16]. The simplest nontrivial CA are binary and one-dimensional, where a cell's neighbours are the cells on either side of it. The name of a CA is a decimal number which, in binary, gives the rule table. According to rule 90, the value of a particular cell i is the sum modulo 2 of the values of its two neighbour cells on the previous time step t . Rule 150 also includes the value of cell i at time step t . These two rules may be defined as $x_i^{t+1} = x_{i-1}^t + x_{i+1}^t$ and $x_i^{t+1} = x_{i-1}^t + x_i^t + x_{i+1}^t$ respectively, where x_i^t represents the state value of cell i at time t .

Null CA are those where cells with permanent null content are supposed adjacent to the extreme cells of the CA. Binary CA where the neighbourhood dependence is just on XOR operations are called linear CA. If in a CA different rules are applied over different cells, then it is called a hybrid CA. Linear Hybrid Cellular Automata are usually denoted with the acronyms LHCA. In this research only one-dimensional 90/150 null LHCA are considered. Binary string $R_1R_2\dots R_n$ are here used to represent n -cell LHCA, where R_i is either 0, if cell i uses rule 90, or 1, if cell i uses rule 150, for $1 \leq i \leq n$.

Given an irreducible polynomial, several algorithms have been developed to find its corresponding LHCA. The most recent one [1] applies the Euclidean algorithm to compute the LHCA in a polynomial running time, so it is sufficiently fast to generate LHCA for polynomials of very large degree. On the other hand, in [3], a synthesis algorithm based also on the Euclidean algorithm which allows to calculate in linear time the characteristic polynomial for any given LHCA was introduced. In this work such an algorithm will be called *Polynomial-Synthesis Algorithm*.

The SG is a well-known keystream generator introduced by Coppersmith, Krawczyk and Mansour in 1993 [4]. It is composed of two LFSRs: A selector register that produces a sequence used to decimate the sequence generated by the other register. The selector register is here denoted by S , its length is L_S ,

its characteristic polynomial is $P_S(x)$ and the sequence it produces is $\{s_i\}$. The decimated sequence is denoted $\{a_i\}$, the second register that produces it is A , its length is L_A , its characteristic polynomial is $P_A(x)$, and the shrunken sequence is $\{z_j\}$. The period of the shrunken sequence is $T = (2^{L_A} - 1)2^{L_S - 1}$ and its linear complexity L is such that $L_A 2^{L_S - 2} < L \leq L_A 2^{L_S - 1}$. Its characteristic polynomial is of the form $P(x)^N$ where $P(x)$ is a L_A -degree primitive polynomial and N is an integer such that $2^{L_S - 2} < N \leq 2^{L_S - 1}$. Despite its simplicity, the SG has remained resistant against efficient cryptanalysis.

3 The Cellular Automata-Based Model for the Shrinking Generator

In this work we consider the linear model of the SG described in [6] in terms of LHCA. The equivalent LHCA obtained for any SG through the algorithm proposed in such a work, and here denoted *CA-Synthesis Algorithm*, are formed by concatenations of basic primitive LHCA and their mirror images, with one or two modifications of rules in each LHCA component [15]. In particular, we have found that the numbers of modifications in the described model are two in all but two concatenated LHCA, and only one modification in the two extreme LHCA.

The output of the *CA-Synthesis Algorithm* is formed by two equivalent LHCA for any SG with selector LFSR of length L_S and decimated LFSR sequence produced by A . The characteristic polynomial of the equivalent LHCA is the same as the one of the original SG, that is to say, $P(x)^N$.

Since the number of concatenations is between $2^{L_S - 2}$ and $2^{L_S - 1}$, and the length of the basic primitive LHCA is L_A , we have that the length of the equivalent LHCA is given by an integer L such that $L_A 2^{L_S - 2} < L \leq L_A 2^{L_S - 1}$. Consequently, in order to generate the whole shrunken sequence in one of the extreme cells of the equivalent LHCA it would be necessary to determine uniquely the initial state of the equivalent LHCA which is able to produce it, and to get this, it would be necessary to intercept L shrunken bits. Consequently, although we have a linear model of the SG, in order to break the SG with it, we need as many intercepted bits as the linear complexity of the SG.

This work provides an efficient way to use the CA-model of the SG in order to guess unseen bits of the shrunken sequence from the interception of a number of bits lower than the linear complexity of the SG.

4 Theoretical Basis

Let $Z = Z^0 = z_0, z_1, z_2, \dots$ be the output sequence of the SG whose characteristic polynomial $P(x)^N \in GF(2)[x]$ has degree L . Let $Z^t = z_t, z_{t+1}, z_{t+2}, \dots$ denote the t -th phase shift of Z . Let $\alpha \in GF(2^{L_A})$ be a root of $P(x)$.

Since the equivalent LHCA may generate the shrunken sequence in any of its cells, given a shrunken sequence $z_0, z_1, z_2, \dots, z_r$, it is always possible to

assume, without loss of generality, that its generation is at the left extreme cell so that $x_1^0 = z_0, x_1^1 = z_1, x_1^2 = z_2, \dots, x_1^r = z_r$. According to this, assuming the knowledge of r bits of the shrunken sequence, we may reconstruct r sub-sequences x_i^t of length $r - i + 1$ corresponding to the rules R_i with $1 < i \leq r$ so that $x_i^t = R_{i-1}(x_{i-1}^t, x_{i-1}^{t+1}, x_{i-2}^t)$. Since rules 90 and 150 are additive and the equivalent LHCA is null boundary, for any rule R_i , the previous expression corresponds to a sum of some elements of the shrunken sequence: $x_i^t = z_{t+k_1} + z_{t+k_2} + \dots + z_{t+k_{r_i}}$, whose sub-indexes correspond to the exponents of the unknown in the characteristic polynomial of the LHCA $R_1 R_2 \dots R_{i-1}$. Consequently, if this sub-sequence $\{x_i^t\}$ of length $r - i + 1$ is used recursively as left extreme sequence of the equivalent LHCA in order to reconstruct in the same way as before, $r - i + 1$ sub-sequences of length $r - 2i + 2$ corresponding to the rules R_i with $1 < i \leq r - i + 1$, we obtain in the same cell i : $z_{t+2k_1} + z_{t+2k_2} + \dots + z_{t+2k_{r_i}}$. In this way, if the hypothesis $z_{t+d} = z_{t+2k_1} + z_{t+2k_2} + \dots + z_{t+2k_{r_i}}$, or equivalently

$$Z^d = Z^{2k_1} + Z^{2k_2} + \dots + Z^{2k_{r_i}} \tag{1}$$

is fulfilled, then a d -th phase shift of the shrunken sequence reappears at cell i of the equivalent LHCA each second chained sub-triangle generated as explained in the previous paragraph.

Furthermore, it is easy to see that if hypothesis (1) is satisfied, then each $2j$ -th chained triangle provides $r - 2ji + 2j$ bits of a jd -th phase shift of the shrunken sequence.

Note that hypothesis (1) may be easily generalized to hypothesis:

$$Z^d = Z^{2^l k_1} + Z^{2^l k_2} + \dots + Z^{2^l k_{r_i}} \tag{2}$$

so that if hypothesis (2) is satisfied, the reappearance of a d -th phase shift of the shrunken sequence is guaranteed at cell i in each 2^l -th chained sub-triangle.

On the other hand, the exponents of the unknown in the characteristic polynomial of the LHCA $R_1 R_2 \dots R_i$ with $i \leq L_A$ are of the form $(2k_1, 2k_2, \dots, 2k_{r_i})$ if and only if the LHCA is the concatenation of a basic automata with its mirror image, $R_1 R_2 \dots R_{i/2} R_{i/2} \dots R_2 R_1$. Also, the exponents of the unknown in the characteristic polynomial of the LHCA $R_1 R_2 \dots R_i$ with $i \leq L_A$ are of the form $(2k_1 + 1, 2k_2 + 1, \dots, 2k_{r_i} + 1)$ if and only if the LHCA is the concatenation of a basic automata with the rule 90 and its mirror image, $R_1 R_2 \dots R_{(i-1)/2} 0 R_{(i-1)/2} \dots R_2 R_1$. Consequently, the hypothesis (2) is always fulfilled for rule R_i when the sub-automata $R_1 R_2 \dots R_i$ corresponds to one of both previous descriptions, and in such a case a d -th phase shift of the shrunken sequence at cell i is obtained in at most the $2^{L_S - 2}$ -th chained sub-triangle.

It is well-known that if $\{s_n\}$ is a sequence produced by a LFSR whose characteristic polynomial is irreducible, and α is a root of such a polynomial, then each element s_n of the sequence may be written as the trace of the n -power of α [9]. Since $P(x)$ is a L_A -degree primitive polynomial, the successive powers $\alpha^i, 0 \leq i < 2^{L_A} - 1$ generate the finite field $GF(2^{L_A})$, and their respective traces equal the corresponding elements s_i of the PN-sequence associated to the polynomial $P(x)$. On the other hand, since the trace function is linear and all the

powers of α may be expressed in terms of the first $L_A - 1$ powers, the association between powers of α and elements of the PN-sequence may be transferred to linear relations between different phase shifts of the PN-sequence and the first $L_A - 1$ phase shifts.

From [6] we know that the shrunken sequence is composed of interpolations of different phase shifts of the PN-sequence associated to the polynomial $P(x)$, so that the element s_i of the basic PN-sequence corresponds to the shrunken bit z_{iN} . Consequently, any linear relation between different phase shifts of the PN-sequence deduced as explained in the previous paragraph corresponds to a linear relation between different phase shifts of the shrunken sequence, which are the same phase shifts obtained for the PN-sequence, but multiplied by N .

5 Cryptanalysis of the Shrinking Generator

Starting from the theoretical basis of the previous section, in the following we describe an efficient algorithm based on the generation of the finite field $GF(2^{L_A})$ and the test of hypothesis (2). The proposed cryptanalysis algorithm has two phases. The first off-line phase is equal for many different generators that have part of its structure in common. The second on-line phase requires as input r intercepted shrunken bits, and provides as output a number of unseen shrunken bits which is directly proportional to the number of intercepted shrunken bits.

Algorithm

Off-line Phase:

Input: The lengths L_S and L_A , and the characteristic polynomial of A , $P_A(x)$ corresponding to the LFSRs S and A components of the SG.

Step 1: Using the *CA-Synthesis Algorithm* described in [6], compute the two equivalent LHCA that are valid for any SG with selector LFSR of length L_S and decimated LFSR sequence produced by A .

Step 2: Using the primitive L_A -degree polynomial $P(x)$ associated to the basic LHCA, generate the finite field $GF(2^{L_A})$ formed with the exponentiation of one root of such a polynomial, α , and express each element of $GF(2^{L_A})$, α^e as a L_A -length array $E = [e_0, e_1, \dots, e_{L_A-2}, e_{L_A-1}]$ where $e_i = 1$ iff α^i is present in the expression of α^e .

Step 3: For both LHCA obtained in step 1, search for sub-automata of the form $R_1R_2 \dots R_{i/2}R_{i/2} \dots R_2R_1$ or $R_1R_2 \dots R_{(i-1)/2}0R_{(i-1)/2} \dots R_2R_1$. For each of them, use the *Polynomial-Synthesis Algorithm* described in [3] to calculate the $2 * r$ characteristic polynomials and express them as $2 * r$ different $(L + 1)$ -length arrays $D = [d_0, d_1, \dots, d_{L-1}, d_L]$ where $d_i = 1$ iff i is an exponent of the unknown in the corresponding polynomial. Using the finite field generated in the previous step, decompose each array D as an equivalent linear expression $a * B + C$ with a being a power of 2, and B and C two arrays such that $B = [0, b_1, \dots, b_{L_A-2}, b_{L_A-1}]$ and $C = [c, c, \dots, c]$. Consider B as the output of this step.

Step 4: For each array $B = [0, b_1, \dots, b_{L_A-2}, b_{L_A-1}]$ obtained in Step 3, search it within all the L_A -length arrays obtained in Step 2, and if found it, associate

to the corresponding rule the $(N/a) * (a * e + c)$ -th phase shift for the N/a -th sub-triangles.

Output: For each equivalent LHCA, favourable rules and corresponding phase shifts and sub-triangles.

On-line Phase:

Input: Output of the off-line phase and r intercepted shrunken bits.

Step 5: Once intercepted r shrunken bits, proceed with them by generating the chained sub-triangles indicated in Step 4, to obtain for all successful rules u , $\lfloor \sum ar^2/2NR_u \rfloor$ bits of the $(N/a) * (a * e + c)$ -th phase shifts of the shrinking sequence.

Output: A variable number of bits of the shrunken sequence.

Note that all the computations made for any LHCA are useful for any other LHCA with the same basic CA and more concatenations, that is to say, the outputs of steps 2, 3 and 4 obtained for an equivalent LHCA with characteristic polynomial $(P(x))^{N_1}$ continue being correct for any other equivalent LHCA with characteristic polynomial $(P(x))^{N_2}$, with $N_1|N_2$. Consequently, the cryptanalysis of a SG with LFSRs S_1 and A_1 are useful for the cryptanalysis of any other SG with LFSRs S_2 and A_2 such that its corresponding characteristic polynomial is $(P_A(x))N_2$ with $N_1|N_2$.

Example:

Consider any SG with $L_S=3$, $L_A=5$, $P_A(x) = 1 + x + x^2 + x^3 + x^5$.

Step 1: $P(x) = 1 + x + x^2 + x^4 + x^5$

The two LHCA computed using the *CA-Synthesis Algorithm* are:

$LHCA_1$:10001100000000110001 and

$LHCA_2$: 00000000011000000000

Step 2: The finite field $GF(2^5)$ associated to $P(x)$ is computed. Here only the powers that have the term 1 in their expressions are showed.

$$\begin{aligned}
 E_5 &: [0124] & E_6 &: [034] & E_7 &: [02] & E_{10} &: [01234] & E_{11} &: [03] & E_{13} &: [014] \\
 E_{14} &: [04] & E_{15} &: [024] & E_{16} &: [0234] & E_{17} &: [023] & E_{19} &: [01] & E_{23} &: [012] \\
 E_{26} &: [0123] & E_{28} &: [013] & E_{30} &: [0134] & E_{31} &: [0]
 \end{aligned}$$

Steps 3 and 4: For $LHCA_1$ the only useful sub-automata corresponds to rule 5. For $LHCA_2$ we find useful sub-automata from rule 2 to rule 9. The characteristic polynomials for those rules are computed with the *Polynomial-Synthesis Algorithm*. Then they are decomposed in terms of 5-length arrays. Here only the positive terms of the arrays are shown and the arrays C are written as c . After the arrows the corresponding phase shifts and sub-triangles obtained from the comparison with the arrays computed in step 2 are indicated.

For $LHCA_1$:

$$D_5^1: [1 \ 3 \ 5]: 2*[0 \ 1 \ 2] + 1 \rightarrow 2(2*23+1) = 94\text{-th shift, 2-nd triangles}$$

For $LHCA_2$:

$$D_5^2: [0 \ 2]: 2*[0 \ 1] \rightarrow 4*19 = 76\text{-th shift, 2-nd triangles}$$

$$D_3^2: [3]: [0] + 3 \rightarrow 3\text{-th shift, 1-st triangle}$$

$$D_4^2: [0 \ 2 \ 4]: 2*[0 \ 1 \ 2] \rightarrow 4*23 = 92\text{-th shift, 2-nd triangles}$$

- $D_5^2: [1\ 5]: 4*[0\ 1]+1 \rightarrow 4*19+1 = 77$ -th shift, 1-st triangle
- $D_6^2: [0\ 4\ 6]: 2*[0\ 2\ 3] \rightarrow 4*17 = 68$ -th shift, 2-nd triangles
- $D_7^2: [7]: [0]+7 \rightarrow 7$ -th shift, 1-th triangle
- $D_8^2: [0\ 4\ 6\ 8]: 2*[0\ 2\ 3\ 4] \rightarrow 4*16 = 64$ -th shift, 2-nd triangles
- $D_9^2: [1\ 5\ 9]: 4*[0\ 1\ 2]+1 \rightarrow 4*23+1 = 93$ -th shift, 1-st triangle

From the off-line phase we know that we may deduce a number of unseen shrunken bits from positions 94, 76, 92, 77, 68, 64 and 93 (mod 124), which depend on the number of intercepted bits in the on-line phase.

Step 5: After the interception of $r=10$ shrunken bits: 0011101011, we proceed with them by generating the four chained sub-triangles corresponding to the $LHCA_2$ for Rule R_2

R_1	R_2	R_3	chain	R_1	R_2	R_3	chain	R_1	R_2	R_3	chain	R_1	R_2	R_3
90	90	90		90	90	90		90	90	90		90	90	90
0	0	1		1	1	1		1	0	0		0	1	1
0	1	1		1	0	0		0	0	1		1	0	1
1	1	0		0	1	0		0	1	0		0	0	
1	1	1		1	0	1		1	0	1		1		
1	0	0		0	0	0		0	0					
0	1	0		0	0	1		1						
1	0	0		0	1									
0	1	1		1										
1	1													
1														

So, we get 6 bits of the 76-th phase shift, $z_{76}:1, z_{77}:0, z_{78}:0, z_{79}:1, z_{80}:0, z_{81}:1$, and 2 bits of the 28-th phase shift, $z_{28}:1, z_{29}:1$. We may also use R_2 of $LHCA_1$ to generate with the 4-th triangle, 2 new bits: $z_{92}:0, z_{93}:1$. In summary, from the knowledge of 10 shrunken bits we have discovered other 10 unseen shrunken bits.

6 Analysis of the Algorithm

In this section we present some simulation results and a comparison with previous attacks. In the following we show a summary of the outputs from the off-line phase of the attack using the first LHCA from the tables in [2]. The next table contains the rules that allow to get new bits from an earlier triangle than 2^{L_S-1} -th.

L_A	6	7	8	10	12	13	15	16	17	20	21	23	25	26	27	29	30
R_i	2	3	6	2;4	4	10	8	2	13	4	6	8	3	6	2	6;13	3
L_A	31	32	33	34	35	36	37	38	39	40	42	43	45	46	47	48	49
R_i	2;4	2;3;4	2	5	2;13	6	5	2;4	5	2;6	2;7	2	2	6	4	5	6

As the lengths of the LFSRs increase, the proposed attack naturally needs a larger intercepted sequence. Concretely, the number r of intercepted bits that are necessary to guess n new bits is given by $R_i * 2^{L_S-2} + n$. Note that the

automata with more favorable rules associated give more information for each intercepted bit. The most interesting conclusion of this empirical analysis is that most basic automata have some sub-automata of the form $R_1R_2 \dots R_{i/2}R_{i/2} \dots R_2R_1$ or $R_1R_2 \dots R_{(i-1)/2}0R_{(i-1)/2} \dots R_2R_1$, which allow to use them in order to get new unseen shrunken bits starting from a number of intercepted shrunken bits that is lower than the linear complexity of the generator.

In order to compare the proposed attack with known related results, two important aspects of our algorithm should be highlighted. First, the off-line phase is to be executed before intercepting sequence, and consequently, its computational complexity should not be considered in the same way as on-line computations (after interception). The second point is the determinism of the proposed attack because the obtained bits are known with absolute certainty. The computational complexity of the proposed attack is $O(2^{L_S-2})$. If we compare it with the one of known attacks on SG, we find the following. The complexity of the divide-and-conquer attack proposed in [14] is exponential in L_S . The probabilistic correlation attack described in [7] has a computational complexity of $2^{L_A} * L_A^2$. Also the probabilistic correlation attack introduced in [8] is exponential in L_A . Finally, the distinguishing attack investigated in [5] needs approximately 2^{32} bits to distinguish the SG with a weight 4 polynomial of degree 10000. However, note that the distinguishing attack does not try to recover the sequence, and instead, the aim is to distinguish the keystream from a purely random sequence.

7 Conclusions and Open Problems

The main purpose of this paper has been to introduce a practical known-plaintext attack on the shrinking generator, which requires a number of intercepted bits lower than the linear complexity in order to predict with absolute certainty approximately a number of unseen shrunken bits that depends on the number of intercepted bits.

Any shrinking generator whose characteristics lead to a successful off-line phase of the algorithm that implies the deduction of too many unseen shrunken bits should be rejected for its cryptographic use. Therefore, the proposed algorithm is useful both for cryptanalysts and for cryptographers who use the shrinking generator. The two main advantages of the described cryptanalysis are the determinism of bits prediction and the application of obtained results to different shrinking generators.

One of the subjects that are being object of work in progress is the modelling of other keystream generators through concatenations of maximum length CA.

References

1. K. Cattell and J.C. Muzio. Synthesis of one-dimensional linear hybrid cellular automata. IEEE Transactions on Computer-Aided Design, 1996.
2. K. Cattell and J.C. Muzio. Tables of linear cellular automata for minimal weight primitive polynomials of degree up to 300. Technical report, University of Victoria, Department of Computer Science, 1993.

3. K. Cattell, S. Zhang, X. Sun, M. Serra, J.C. Muzio, and D. M. Miller, One-Dimensional Linear Hybrid Cellular Automata: Their Synthesis, Properties, and Applications in VLSI Testing. Tutorial. www.cs.uvic.ca/~mserra/CA.html
4. D. Coppersmith, H. Krawczyk, Y. Mansour, The Shrinking Generator. Proc. Crypto'93. LNCS 773, Springer Verlag (1994) 22-39.
5. P. Ekdahl, W. Meier, T. Johansson, Predicting the Shrinking Generator with Fixed Connections. Proc. Eurocrypt'03. LNCS 2656, Springer Verlag (2004) 345-359.
6. A. Fúster-Sabater, D. de la Guía, Cellular Automata Application to the Linearization of Stream Cipher Generators. Proc. ACRI 2004. LNCS 3305, Springer Verlag (2004) 612-622.
7. J.D.Golic, L O'Connor, A Cryptanalysis of Clock-Controlled Shift Registers with Multiple Steps. Cryptography: Policy and Algorithms (1995) 174-185.
8. T. Johansson, Reduced Complexity Correlation Attacks on Two Clock-Controlled Generators. Proc. Asiacrypt'98. LNCS 1514, Springer Verlag (1998) 342-356.
9. E.L. Key, An Analysis of the Structure and Complexity of Nonlinear Binary Sequence Generators, IEEE Transactions on Information Theory 22 (1976) 732-736.
10. W.Meier, O.Staffelbach, Analysis of Pseudo Random Sequence Generated by Cellular Automata. Proc. Eurocrypt'91. LNCS 547, Springer Verlag (1992) 186-199.
11. M. Mihaljevic, Security Examination of a Cellular Automata Based Pseudorandom Bit Generator Using an Algebraic Replica Approach. Proc. AAECC-12. LNCS 1255, Springer Verlag (1997) 250-262.
12. S.Nandi, B.K.Kar, P.P.Chaudhuri, Theory and Applications of Cellular Automata in Cryptography. IEEE Transactions on Computers 43,12 (1994) 1346-1357.
13. F. Serebinski, P. Bouvry, A.Y. Zomaya, Cellular Automata Computations and Secret Key Cryptography. Parallel Computing archive Volume 30, Issue 5-6 (2004).
14. L.Simpson, J.D.Golic, E.Dawson, A Probabilistic Correlation Attack on the Shrinking Generator. Proc.ACISP'98. LNCS 1438, Springer Verlag (1998) 147-158.
15. X. Sun, E. Kontopidi, M. Serra, and J. C. Muzio. The concatenation and partitioning of linear finite state machines. International Journal of Electronics, 78(5) 809-839, 1995.
16. S.Wolfram, Cryptography with Cellular Automata. Proc. Crypto'85. LNCS 218, Springer Verlag (1986) 429-432.

A Comparative Study of Proposals for Establishing Security Requirements for the Development of Secure Information Systems

Daniel Mellado¹, Eduardo Fernández-Medina², and Mario Piattini²

¹ Ministry of Labour and Social Affairs, Management Organism of Information Technologies of the Social Security, Quality, Auditing and Security Institute, Madrid, Spain

Daniel.Mellado@alu.uclm.es

² Alarcos Research Group, Information Systems and Technologies Department, UCLM-Soluziona Research and Development Institute, University of Castilla-La Mancha, Paseo de la Universidad 4, 13071 Ciudad Real, Spain
(Eduardo.FdezMedina, Mario.Piattini)@uclm.es

Abstract. Nowadays, security solutions are focused mainly on providing security defences, instead of solving one of the main reasons for security problems that refers to an appropriate Information Systems (IS) design. In this paper a comparative analysis of eight different relevant technical proposals, which place great importance on the establishing of security requirements in the development of IS, is carried out. And they provide some significant contributions in aspects related to security. These can serve as a basis for new methodologies or as extensions to existing ones. Nevertheless, they only satisfy partly the necessary criteria for the establishment of security requirements, with guarantees and integration in the development of IS. Thus we conclude that they are not specific enough for dealing with security requirements in the first stages of software development in a systematic and intuitive way, though parts of the proposals, if taken as complementary measures, can be used in that manner.

1 Introduction

Present-day information systems are vulnerable to a host of threats. What is more, with increasing complexity of applications and services, there is a correspondingly greater chance of suffering from breaches in security [25]. In our contemporary Information Society, depending as it does on a huge number of software systems which have a critical role, it is absolutely vital that IS are ensured as being safe right from the very beginning [2, 18]. That is so, is obvious from the potential losses faced by organizations that put their trust in all these IS.

As we know, the principle which establishes that the building of security into the early stages of the development process is cost-effective and also brings about more robust designs is widely-accepted [15]. The biggest problem, however, is that in the majority of software projects security is dealt with when the system has already been designed and put into operation. On many occasions, this is thanks to an inappropriate management of the specification of the security requirements of the new system, since the stage known as the requirement specification phase is often carried out with the

aid of just a few descriptions, or the specification of objectives that are put down on a few sheets of paper. Added to this, the actual security requirements themselves are often not well understood. This being so, even when there is an attempt to define security requirements, many developers tend to describe design solutions in terms of protection mechanisms, instead of making declarative propositions regarding the level of protection required [8].

A very important part of the achieving of secure software systems in the software development process is that known as Security Requirements Engineering. This provides techniques, methods and norms for tackling this task in the IS development cycle. It should involve the use of repeatable and systematic procedures in an effort to ensure that the set of requirements obtained is complete, consistent and easy to understand and analyzable by the different actors involved in the development of the system [16]. A good requirement specification document should include both functional requirements (related to the services which the software or system should provide), and non-functional (related to what are known as features of quality, performance, portability, security, etc). As far as security is concerned, it should be a consideration throughout the whole development process, and it ought to be defined in conjunction with the requirements specification [19].

In this paper eight relevant technical proposals are studied. They are ones which place importance on eliciting security requirements in the development of IS. These proposals will be presented briefly and then compared in this paper. This should serve as an introduction to the current state of the art of security requirements in the development of IS. The remainder of the paper is set out as follows: in section 2, we will describe each one of the technical proposals. We will present the comparative study performed on these proposals in section 3. Lastly, our conclusions are set out in section 4.

2 Technical Proposals Which Support Security Requirements

The proposals which will be analyzed in this comparative study are as follows:

- Breu, et al. 2004 & Breu & Innerhofer–Oberperfler, 2005: “Towards a systematic development of secure systems” [5] and “Model based business driven IT security analysis” [6].
- Firesmith, 2003 & 2004: “Security Use Cases” [9] and “Security Requirements in Open Process Framework” [10].
- Jennex 2005: “Modeling security requirements for information system development” [13].
- Myagmar, Lee, & Yurcik, 2005: “Threat modeling as a basis for security requirements” [20].
- Toval et al. 2001: “Security requirements in SIREN” [24] and Gutiérrez, et al. 2005: “Security Requirements for Web Services based on SIREN” [11].
- Peeters 2005: “Agile security requirements engineering” [21].
- Popp et al. 2003: “Security-Critical system development with extended use cases” [22].
- Yu 1997: “Security requirements based on the i* framework” [26].

We have chosen these proposals because the majority of them try to solve the problem of security in the different phases of IS development. They also place an

emphasis on security requirements in the development of secure Information Systems. We give a brief outline of each of these approaches below.

2.1 Towards a Systematic Development of Secure Systems, and Model Based Business Driven IT Security Analysis (Proposed by Breu et al. 2004 [5] and Breu & Innerhofer–Oberperfler, 2005 [6])

The authors propose a new process model for security engineering, which extends object oriented, use case driven software development by the systematic treatment of security related issues. They also introduce the notion of security aspects, describing the most relevant security requirements and the countermeasures at a certain level of abstraction. Starting from the concept of iterative software construction, they present a micro-process for the security analysis, made up of five steps, which are performed repeatedly at each level of abstraction throughout the incremental development: elicitation of security requirements, threats and risks analysis, taking measures and the correctness check relating measures and requirements. Finally, the authors conclude that security of information is a business issue, and for this reason its management should be business driven.

2.2 Security Use Cases, and Security Requirements in Open Process Framework (Proposed by Firesmith, 2003 [9] & 2004 [10])

Firesmith in [10] offers some steps which allow security requirements to be defined from reusable templates. His analysis of security requirements is founded on two basic principles obtained from OCTAVE (Operationally Critical Threat, Asset, and Vulnerability Evaluation) [1] based on resources and risk-driven. The steps in his process for the identification and analysis of security requirements are: identification of assets; identifying the most likely attackers types; identification of the possible threats to these assets; determining the negative impacts for each vulnerable resource; estimating and prioritizing security risks with respect to vulnerable resources and according to the most relevant threats and their potentially negative impact; choosing security subfactors, to limit the risk to an acceptable level; choosing the relevant templates for each subfactor and security risk; identify the relevant functional requirements; determine the security criteria; define the security metric, along with the minimum level that is acceptable; specify the requirement.

Moreover, the author proposes security use cases as a technique that should be used to specify the security requirements that the application shall successfully protect itself from its relevant security threats [9].

The final suggestions from this author are that, given that systems usually have similar security requirements, templates should be employed to specify the security requirements in such a way that they can easily be re-used from one system to another.

2.3 Modeling Security Requirements for Information System Development (Proposed by Jennex, 2005) [13]

Jennex puts forward the idea of using barrier analysis and the concept of defence in-depth to modify Siponen and Baskervilles's integrated design paradigm [23] into a more graphical and easier to understand and use methodology.

The methodology suggested by the author, then, proposes using barrier analysis diagrams as a graphical method of identifying and documenting security requirements. Furthermore, this approach used meta-notation to add security details to existing system development diagrams. The process follows the approach of integrating security design into the software development life-cycle. The objective of using barrier diagrams in the requirement phase, therefore, is that the security requirements should be appropriately identified.

2.4 Threat Modeling as a Basis for Security Requirements (Proposed by Myagmar, Lee, & Yurcik, 2005) [20]

The authors take as starting point for their proposal the following question, one that is important to ask in every IS- “Are the system’s security features really necessary, and do they really meet the system’s security needs?”

The writers offer a viewpoint on the process of requirement engineering in which, with an appropriate identification of threats and a proper choice of countermeasures, the ability of attackers to misuse or abuse the system is lessened.

The threat-modelling process set out by these authors is made up of three high-level steps: Characterizing the system; Identifying assets and access points; Identifying threats.

As far as the specification of security requirements is concerned, the greater part of the information needed for the elicitation of requirements and for composing an initial set of security requirements is provided by means of threat modelling. This is done by changing a declaration of threat into a requirement by including “must not” in the declaration.

Lastly, with the final goal being to achieve 100% risk acceptance, the risk management the writers propose consists of: risk assessment (to do this risks should be prioritized according to the damage they might cause and to the likelihood of their occurring), risk reduction, and risk acceptance.

2.5 Security Requirements in SIREN and Security Requirements for Web Services Based on SIREN (Proposed by Toval, et al. 2001 [24] and Gutierrez, et al. 2005 [11])

In [24] Toval et al. define a Requirement Engineering process, based on the re-use of security requirements, which is also compatible with MAGERIT (the Spanish public administration risk analysis and management method), which conforms to CCF (Common Criteria Framework) defined by the ISO 15408 (ISO/IEC, 1999). The re-use of security requirements is carried out at different specification level: at a documentation level through the defining of a hierarchical structure of security requirement specifications, and at the level of security requirement by means of its being stored in the repository of re-usable requirements. SIREN (Simple REuse of software requireMnts) describes a process model, some basic guidelines, techniques and tools. The guidelines consist of a hierarchy of requirement specification documents, together with the template for each document. It is a spiral model process, and includes the phases of requirements elicitation, requirements analysis and negotiation, requirements specification and validation. A repository of requirements classified by domains and profiles is also defined.

Moreover, in [11] Gutiérrez et al. present a catalogue of security requirement templates for Web Services (WS) based on the SIREN method of requirement engineering. They focus their efforts on the security requirement templates for the following subfactors: authentication, authorization, confidentiality, integrity and privacy.

2.6 Agile Security Requirements Engineering (Proposed by Peeters, 2005) [21]

Peeters proposes to extend agile practices to deal with security in an informal, communicative and assurance-driven spirit. To increase the agility of requirement engineering, Peeters puts forward the idea of using “abuser stories”. These identify how the attackers may abuse the system and jeopardize stakeholders’ assets. Thus, throughout the abuser stories, the establishing of security requirements is made easier. As with “user stories”, “abuser stories” are short and informal and they are scored and ranked according to the perceived threat they pose to customers’ assets. Correct planning will consequently mean considering the “user stories” and the “abuser stories” together. This will ensure an explicit, rational trade-off between functionality and security.

2.7 Security-Critical System Development with Extended Use Cases (Proposed by Popp, et al. 2003) [22]

What these authors provide is an extension to the conventional process of developing use-case-oriented processes [7, 12]. This process normally consists of three activities as far as requirement engineering is concerned.

1. They deal with the static concepts of the domain of an application in a class model known as Application Core. In this point they extend the domain by modelling access policies and security properties based on UMLSec [14].
2. Identification of the use cases and their manifestation in a Use Case Model. These are completed by the textual description coming from characteristics which measure the threats and vulnerability of input and output. They also outline the security policies which are a response to previous threats. The outcome is a model known as Model of Security Use Cases.
3. Integration of the previous two viewpoints in a single oriented object model, mainly through the description of use cases in terms of message flows between objects. The extension consists of the integration of the Security Use Case Model, and the Application Core refines the security policy in terms of the message flows between objects.

2.8 Security Requirements Based on the *i Framework (Proposed by Yu, 1997) [26]**

The *i** framework provides a framework which allows the easy integration of different techniques and concepts for dealing with Systems Security. The structural representation defined in *i** shows the dependence relationships between the actors, and is what makes security aspects appear.

In (Fig. 1. Requirement elicitation process with i^*) we see the process of functional requirements elicitation and analysis defined by i^* , and how it is integrated into the process of elicitation and analysis of security requirements [17].

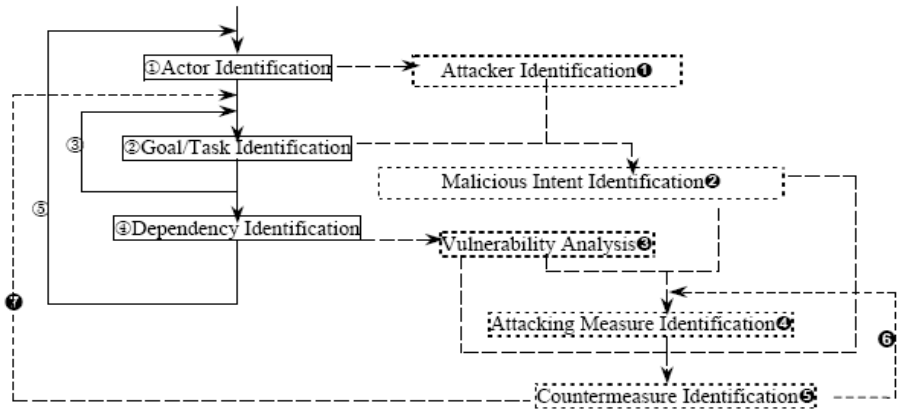


Fig. 1. Requirement elicitation process with i^*

3 Comparison

To get a general overview of the different proposals which we have discussed above, a comparative study of these will be carried out in this section. To this end, we propose an analytical framework based on the following criteria:

- Degree of Agility. This refers to the degree of agility of the methodology of development, as compared with traditional, planned methodologies. To see this we will take as our basis the observations carried out Boehm and Turner [3, 4]. These authors propose a method based on risks, by means of which they try to keep both kinds of methodologies (those which are agile and those driven by planning- traditional ones) in balance, taking advantage of the positive points of the two types and taking steps to make up for their disadvantages. Each proposal will be given a rating using the following scale: high, medium-high, medium, medium-low, low.
- Support. This refers to aspects such as tools, procedures, guides, standards and study cases which help make the proposal easier to use. Each proposal will be given a rating using the following scale: high, medium-high, medium, medium-low, low.
- Degree of integration with other software requirements. This is all about how the establishing of security requirements fits in with the establishing of other software requirements (with the other non-functional requirements, as well as with the functional ones) in the development of an IS. To do this it will take into account aspects such as the use of similar and already-existing techniques for the determining of other requirements, such as, for instance, UML diagrams. It also bears in mind the degree of parallelization and co-ordination with the elicitation of other requirements, etc. Each proposal will be given a rating using the following scale: high, medium-high, medium, medium-low, low.

- User friendliness. Here the reference is to the ease with which the technique could be used without any previous knowledge or special training. In this case characteristics such as the help support and the use of techniques which already exist for other requirements will be taken into account, as well as widely-used standards, etc. Each proposal will be given a rating using the following scale: high, medium-high, medium, medium-low, low.
- Contributions of the proposal as regards security. New perspectives which the proposal brings to the improvement of the establishing of security requirements.

In the following table (Table 1. Comparison of proposals) the comparison between the proposals from different authors, within the analysis framework we propose, is set out.

Table 1. Comparison of proposals

Criteria Proposals	Degree of Agility	Help Support	Degree of integration with other software requirements	User friendli- ness	Contributions of the proposal as regards security
Breu, et al. 2004 [5] and Breu & Innerhofer-Oberperfler, 2005 [6]	Low	Medium	Medium	Medium-High	<ul style="list-style-type: none"> ▪ A micro-process for the security analysis
Firesmith, 2003 and 2004 [9, 10]	Medium	Medium-High	Medium	Medium-High	<ul style="list-style-type: none"> ▪ Security use cases. ▪ Re-usable templates
Jennex, 2005[13]	High	Medium-Low	Medium-High	High	<ul style="list-style-type: none"> ▪ Diagrams of barriers
Myagmar, Lee, & Yurcik 2005 [20]	Medium-Low	Medium	Medium	Medium-High	<ul style="list-style-type: none"> ▪ Threat modeling as a basis for security requirements
Peeters, 2005[21]	High	Medium-Low	High	Medium-High	<ul style="list-style-type: none"> ▪ Abuser stories
Popp, et al. 2003 [22]	Medium-Low	High	Medium-High	Medium-High	<ul style="list-style-type: none"> ▪ UMLSec
Toval,, et al. 2001 [24] & Gutierrez, et al.2005 [11]	Low	Medium-High	Medium	Medium-High	<ul style="list-style-type: none"> ▪ Re-use of security requirements compatible with MAGERIT ▪ A catalogue of security requirements templates for WS
Yu, 1997 [26]	Low	Medium-Low	Medium-High	Medium-High	<ul style="list-style-type: none"> ▪ Integration of functional and security requirements

As can be seen in the table, after our analysis we reach the conclusion that the proposals discussed above present some weaknesses. These include the difficulty of integrating them into the software development; the lack of an overall/complete support of security modelling at an organizational, conceptual and technical level. There is also an increasing distance between the development of the IS and the implementation of the necessary security. Moreover, these proposals are not specific enough for a systematic and intuitive treatment of IS security requirements in the first stages of software development. In short, the proposals we have analyzed partially satisfy the criteria that are necessary for the establishing of security requirements with some degree of guarantee. They do not reach the desired level of integration in the development of IS. At the same time, having said all that, each one of these methodologies contributes highly important aspects to do with security. These are features that can be used as the basis for new methodologies, or as extensions of those that already exist.

4 Conclusions

In our present so-called Information Society the increasingly crucial nature of IS with corresponding levels of new legal and governmental requirements is obvious. So the development of more and more sophisticated approaches to ensuring the security of information is becoming a necessity. Information Security is usually only tackled from a technical viewpoint at the implementation stage, even though it is an important aspect. We believe it is fundamental to deal with security at all stages of IS development, especially in the establishing of security requirements, since these form the basis for the achieving of a robust IS. Various interesting methodological proposals which have to do with this issue exist some of them have been described and compared in this paper, although they all present some weak points, as we have said above. In a similar vein, it must be said that these approaches are not specific enough for a treatment of IS security requirements in the first stages of the IS development process.

Consequently, we consider that it would be interesting to obtain some systematic and intuitive way of eliciting and defining security requirements with some guarantee. Such a technique should allow integration of security requirements into the IS development as far and as much as is possible. It will also permit the re-use of requirements from some projects to others. Added to these considerations is the fact that it will have to be valid for the new Internet-based IS and especially for those based on SOA architecture, supported by the technology of Web Services. To this end it would be good if it provided support tools. Positive it would also be if it were based on standards of normalization for the definition of requirements. Some examples of this might be XML or the use of templates; modeling standards like UML can be used to similar advantage. It would also be desirable for it to conform to security management standards such as ISO/IEC 17799 or COBIT.

Acknowledgements

This paper has been produced in the context of the DIMENSIONS (PBC-05-012-2) Project of the Consejería de Ciencia y Tecnología de la Junta de Comunidades de Castilla-La Mancha along with FEDER and the CALIPO (TIC2003-07804-CO5-03)

and RETISTIC (TIC2002-12487-E) projects of the Dirección General de Investigación del Ministerio de Ciencia y Tecnología.

References

1. Alberts, C.J., Behrens, S.G., Pethia, R.D., and Wilson, W.R., *OCTAVE Framework, Version 1.0*. 1999: Networked Systems Survivability Program. p. 84.
2. Baskeville, R., *The development duality of information systems security*. Journal of Management Systems, 1992. **4**(1): p. 1-12.
3. Boehm, B. and Turner, R., *Observations on Balancing Discipline and Agility*. 2003: Agile Development Conference (ADC '03). p. 32.
4. Boehm, B. and Turner, R., *Balancing Agility and Discipline: Evaluating and Integrating Agile and Plan-Driven Methods*. 2004: ICSE'04. p. 718-719.
5. Breu, R., Burger, K., Hafner, M., and Popp, G., *Towards a Systematic Development of Secure Systems*. 2004: WOSIS 2004.
6. Breu, R. and Innerhofer-Oberperfler, F., *Model based business driven IT security analysis*. 2005: SREIS 2005.
7. D'Souza, D.F. and Wills, A.C., *Objects, Components & Frameworks with UML: The Catalysis Approach*. 1998: Addison-Wesley Publishing Company.
8. Firesmith, D.G., *Engineering Security Requirements*. Journal of Object Technology, 2003. **2**(1): p. 53-68.
9. Firesmith, D.G., *Security Use Cases*. 2003: Journal of Object Technology. p. 53-64.
10. Firesmith, D.G., *Specifying Reusable Security Requirements*. 2004: Journal of Object Technology. p. 61-75.
11. Gutiérrez, C., Moros, B., Toval, A., Fernández-Medina, E., and Piattini, M., *Security Requirements for Web Services based on SIREN*. Symposium on Requirements Engineering for Information Security (SREIS-2005), together with the 13th IEEE International Requirements Engineering Conference – RE'05, 2005.
12. Jacobson, I., Booch, G., and Rumbaugh, J., *The Unified Software Development Process*. 1999: Addison-Wesley Longman Inc.
13. Jennex, M.E., *Modeling security requirements for information systems development*. 2005: SREIS 2005.
14. Jürjens, J., *Secure Systems Development with UML*. 2005: Springer. 309.
15. Kim, H.-K., *Automatic Translation From Requirements Model into Use Cases Modeling on UML*, C. Youn-Ky, Editor. 2005: ICCSA 2005.
16. Kotonya, G. and Sommerville, I., *Requirements Engineering Process and Techniques*. 1998.
17. Liu, L., Yu, E., and Mylopoulos, J., *Security and Privacy Requirements Analysis within Social Setting*. 2003: 11th IEEE International Requirements Engineering Conference.
18. McDermott, J. and Fox, C. *Using Abuse Case Models for Security Requirements Analysis*. in *Annual Computer Security Applications Conference*. 1999. Phoenix, Arizona.
19. Mouratidis, H., Giorgini, P., Manson, G., and Philp, I. *A Natural Extension of Tropos Methodology for Modelling Security*. in *Workshop on Agent-oriented methodologies, at OOPSLA 2002*. 2003. Seattle, WA, USA.
20. Myagmar, S., J. Lee, A., and Yurcik, W., *Threat Modeling as a Basis for Security Requirements*. 2005: SREIS 2005.
21. Peeters, J., *Agile Security Requirements Engineering*. 2005: SREIS 2005.

22. Popp, G., Jürjens, J., Wimmel, G., and Breu, R., *Security-Critical System Development with Extended Use Cases*. 2003: 10th Asia-Pacific Software Engineering Conference. p. 478-487.
23. Siponen, M. and Baskerville, R., *A new paradigm for adding security into IS development methods*. 2001: 8th Annual Working Conference on Information Security Management and Small Systems Security.
24. Toval, A., Nicolás, J., Moros, B., and García, F., *Requirements Reuse for Improving Information Systems Security: A Practitioner's Approach*. 2001: Requirements Engineering Journal. p. 205-219.
25. Walton, J.P., *Developing a Enterprise Information Security Policy*. 2002, ACM Press: Proceedings of the 30th annual ACM SIGUCCS conference on User services.
26. Yu, E., *Towards Modelling and Reasoning Support for Early-Phase Requirements Engineering*. 1997: 3rd IEEE International Symposium on Requirements Engineering (RE'97). p. 226-235.

Stochastic Simulation Method for the Term Structure Models with Jump

Kisoeb Park¹, Moonseong Kim², and Seki Kim^{1,*}

¹ Department of Mathematics, Sungkyunkwan University,
440-746, Suwon, Korea
Tel.: +82-31-290-7030, 7034
{kisoeb, skim}@skku.edu

² School of Information and Communication Engineering,
Sungkyunkwan University, 440-746, Suwon, Korea
Tel.: +82-31-290-7226
moonseong@ece.skku.ac.kr

Abstract. Monte Carlo Method as a stochastic simulation method is used to evaluate many financial derivatives by financial engineers. Monte Carlo simulation is harder and more difficult to implement and analyse in many fields than other numerical methods. In this paper, we derive term structure models with jump and perform Monte Carlo simulations for them. We also make a comparison between the term structure models of interest rates with jump and HJM models based on jump. Bond pricing with Monte Carlo simulation is investigated for the term structure models with jump.

1 Introduction

Before mentioning the procedure in derivation of bond pricing models with jumps, we discuss general models of the term structure of interest rates. Approaches to modeling the term structure of interest rates in continuous time may be broadly described in terms of either the equilibrium approach or the no-arbitrage approach even though some early models include concepts from both approaches.

We introduce one-state variable model of Vasicek (1977)[16], Cox, Ingersoll, and Ross (CIR)[5], the extended model of the Hull and White[10], and the development of the model is the jump-diffusion model of the Ahn and Thompson[1] and the Baz and Das[2]. Conventionally, financial variables such as stock prices, foreign exchange rates, and interest rates are assumed to follow a diffusion processes with continuous paths when pricing financial assets. Also, Heath, Jarrow and Morton(HJM)[6] is widely accepted as the most general methodology for term structure of interest rate models.

In pricing and hedging with financial derivatives, jump-diffusion models are particularly important, since ignoring jumps in financial prices will cause pricing and hedging rates. Term structure model solutions under jump-diffusions

* Corresponding author.

are justified because movements in interest rates display both continuous and discontinuous behavior. These jumps are caused by several market phenomena money market interventions by the Fed, news surprise, and shocks in the foreign exchange markets, and so on.

We study a solution of the bond pricing for the term structure models with jump. The term structure models with jump which allows the short term interest rate, the forward rate, the follow a random walk. We compare between the term structure model of interest rate with jump and the HJM model based on jump. We introduce the Monte Carlo simulation. One of the many uses of Monte Carlo simulation by financial engineers is to place a value on financial derivatives. Interest in use of Monte Carlo simulation for bond pricing is increasing because of the flexibility of the methods in handling complex financial instruments. One measure of the sharpness of the point estimate of the mean is Mean Standard Error(MSE). For the term structure models with jump, we study bond prices by the Monte Carlo simulation. Numerical methods that are known as Monte Carlo methods can be loosely described as statistical simulation methods, where statistical simulation is defined in quite general terms to be any method that utilizes sequences of random numbers to perform the simulation.

The structure of the remainder of this paper is as follows. In Section 2, the basic of bond prices with jump are introduced. In Section 3, the term structure models in jump are presented. In Section 4, we calculate numerical solutions using Monte Carlo simulation for the term structure models with jump. In Section 5, we investigate bond prices given for the eight models using the Vasicek and CIR models. Conclusions are in Section 6.

2 Preliminaries for the Bond Prices

2.1 Stochastic Differential Equation with Jump

We will first recall some notations. All our models will be set up in a given complete probability space $(\Omega, \mathcal{F}_t, P)$ and an argued filtration $(\mathcal{F}_t)_{t \geq 0}$ generated by a Wiener process $W(t)$ in \mathcal{R} . We will ignore taxes and transaction costs. We denote by $V(r, r, T)$ the price at time t of a **discount bond**. It follows immediately that $V(r, T, T) = 1$. Now consider a quite different type of random environment. Suppose $\pi(t)$ represents the total number of extreme shocks that occur in a financial market until time t . The time dependence can arise from the cyclical nature of the economy, expectations concerning the future impact of monetary policies, and expected trends in other macroeconomic variables.

In the same way that a model for the asset price is proposed as a lognormal random walk, let us suppose that the interest rate r and the forward rate is governed by a **Stochastic differential equation(SDE)** of the form

$$dr = u(r, t)dt + \omega(r, t)dW^Q + Jd\pi \quad (1)$$

and

$$df(t, T) = \mu_f(t, T)dt + \sigma_f(t, T)dW^Q(t) + Jd\pi. \quad (2)$$

where $\omega(r, t)$ is the instantaneous volatility, $u(r, t)$ is the instantaneous drift, $\mu_f(t, T)$ represents drift function, $\sigma^2_{f_i}(t, T)$ is volatility coefficients, and jump size J is normal variable with mean μ and standard deviation γ .

2.2 The Zero-Coupon Bond Pricing Equation

When interest rates follow the SDE(1), a bond has a price of the form $V(r, t)$; the dependence on T will only be made explicit when necessary. Pricing a bond is technically harder than pricing an option, since there is no underlying asset with which to hedge. We set up a portfolio containing two bonds with different maturities T_1, T_2 . The bond with maturity T_1 has price V_1 , and the bond with maturity T_2 has price V_2 . Thus, the riskless portfolio is

$$\Pi = V_1 - \Delta V_2. \tag{3}$$

And then we applied the jump-diffusion version of Ito's lemma. Hence we derive the partial differential bond pricing equation.

Theorem 1. *If r satisfy Stochastic differential equation $dr = u(r, t)dt + \omega(r, t)dW^Q + Jd\pi$ then the zero-coupon bond pricing equation in jumps is*

$$\frac{\partial V}{\partial t} + \frac{1}{2}\omega^2 \frac{\partial^2 V}{\partial r^2} + (u - \lambda\omega) \frac{\partial V}{\partial r} - rV + hE[V(r + J, t) - V(r, t)] = 0 \tag{4}$$

where $\lambda(r, t)$ is the market price of risk. The final condition corresponds to the payoff on maturity and so $V(r, T, T) = 1$. Boundary conditions depend on the form of $u(r, t)$ and $\omega(r, t)$.

3 Bond Pricing Models with Jump

The dependence of the yield curve on the time to maturity, $T - t$, is called the **term structure of interest rates**. It is common experience from market data that yield curve typically come in three distinct shapes, each associated with different economic conditions. A wide variety of yield curves can be predicted by the model, including **Increasing**, **decreasing**, and **humped**. Now we consider the term structure models with jump.

3.1 Jump-Diffusion Version of Extended Vasicek's Model

The time dependence can arise from the cyclical nature of the economy, expectations concerning the future impact of monetary policies, and expected trends in other macroeconomic variables. In this study, we extend the jump-diffusion version of equilibrium single factor model to reflect this time dependence. We proposed the mean reverting process for interest rate r is given by

$$dr(t) = [\theta(t) - a(t)r(t)]dt + \sigma(t)dW^Q(t) + Jd\pi(t) \tag{5}$$

We will assume that the market price of interest rate diffusion risk is a function of time, $\lambda(t)$. Let us assume that jump risk is diversifiable. From equation (4)

with the drift coefficient $u(r, t) = \theta(t) - a(t)r(t)$ and the volatility coefficient $\omega(r, t) = \sigma(t)$, we get the partial differential difference bond pricing equation:

$$[\theta(t) - a(t)r(t) - \lambda(t)\sigma(t)]V_r + V_t + \frac{1}{2}\sigma(t)^2V_{rr} - rV + hV[-\mu A(t, T) + \frac{1}{2}(\gamma^2 + \mu^2)A(t, T)^2] = 0. \tag{6}$$

Then the yield on zero-coupon bond price expiring $T - t$ periods hence is given by:

$$Y(r, t, T) = -\frac{\ln V(r, t, T)}{T - t} \tag{7}$$

is defined the entries **term structure of interest rates**. The price of a discount bond that pays off \$ 1 at time T is the solution to (6) that satisfies the boundary condition $V(r, T, T) = 1$. A solution of the form:

$$V(r, t, T) = \exp[-A(t, T)r + B(t, T)] \tag{8}$$

can be guessed. Bond price derivatives can be calculated from (7). We omit the details, but the substitution of this derivatives into (6) and equating powers of r yields the following equations for A and B .

Theorem 2

$$-\frac{\partial A}{\partial t} + a(t)A - 1 = 0 \tag{9}$$

and

$$\frac{\partial B}{\partial t} - \phi(t)A + \frac{1}{2}\sigma(t)^2A^2 + h[-\mu A + \frac{1}{2}(\gamma^2 + \mu^2)A^2] = 0, \tag{10}$$

where, $\phi(t) = \theta(t) - \lambda(t)\sigma(t)$ and all coefficients is constants. In order to satisfy the final data that $V(r, T, T) = 1$ we must have $A(T, T) = 0$ and $B(T, T) = 0$.

3.2 Jump-Diffusion Version of Extended CIR Model

We proposed the mean reverting process for interest rate r is given by

$$dr(t) = [\theta(t) - a(t)r(t)]dt + \sigma(t)\sqrt{r(t)}dW^Q(t) + Jd\pi(t) \tag{11}$$

We will assume that the market price of interest rate diffusion risk is a function of time, $\lambda(t)\sqrt{r(t)}$. Let us assume that jump risk is diversifiable.

In jump-diffusion version of extended Vasicek’s model the short-term interest rate, r , to be negative. If Jump-diffusion version of extended CIR model is proposed, then rates are always non-negative. This has the same mean-reverting drift as jump-diffusion version of extended Vasicek’s model, but the standard deviation is proportional to $\sqrt{r(t)}$. This means that its standard deviation increases

when the short-term interest rate increases. From equation (4) with the drift coefficient $u(r, t) = \theta(t) - a(t)r(t)$ and the volatility coefficient $\omega(r, t) = \sigma(t)\sqrt{r(t)}$, we get the partial differential bond pricing equation:

$$[\theta(t) - a(t)r(t) - \lambda(t)\sigma(t)r(t)]V_r + V_t + \frac{1}{2}\sigma(t)^2r(t)V_{rr} - rV + hV[-\mu A(t, T) + \frac{1}{2}(\gamma^2 + \mu^2)A(t, T)^2] = 0. \tag{12}$$

Bond price partial derivatives can be calculated from (12). We omit the details, but the substitution of this derivatives into (6) and equating powers of r yields the following equations for A and B .

Theorem 3

$$-\frac{\partial A}{\partial t} + \psi(t)A + \frac{1}{2}\sigma(t)^2A^2 - 1 = 0 \tag{13}$$

and

$$\frac{\partial B}{\partial t} - (\theta(t) + h\mu)A + \frac{1}{2}h[(\gamma^2 + \mu^2)A^2] = 0, \tag{14}$$

where, $\psi(t) = a(t) + \lambda(t)\sigma(t)$ and all coefficients is constants. In order to satisfy the final data that $V(r, T, T) = 1$ we must have $A(T, T) = 0$ and $B(T, T) = 0$.

Proof. In equations (12) and (13), by using the solution of this Ricatti’s equation formula we have

$$A(t, T) = \frac{2(e^{\omega(t)(T-t)} - 1)}{(\omega(t) + \psi(t))(e^{\omega(t)(T-t)} - 1) + 2\omega(t)} \tag{15}$$

with $\omega(t) = \sqrt{\psi(t)^2 + 2\sigma(t)}$. Similarly way, we have

$$B(t, T) = \int_t^T \left\{ -(\theta(t) + h\mu)A + \frac{1}{2}h(\gamma^2 + \mu^2)A^2 \right\} dt . \tag{16}$$

These equation yields the exact bond prices in the problem at hand. Equation (16) can be solved numerically for B . Since (15) gives the value for A , bond prices immediately follow from equation (6).

3.3 Heath-Jarrow-Merton(HJM) Model with Jump

The HJM consider forward rates rather than bond prices as their basic building blocks. Although their model is not explicitly derived in an equilibrium model, the HJM model is a model that explains the whole term structure dynamics in a no-arbitrage model in the spirit of Harrison and Kreps[?], and it is fully compatible with an equilibrium model. If there is one jump during the period $[t, t + dt]$ then $d\pi(t) = 1$, and $d\pi(t) = 0$ represents no jump during that period.

We know that the **zero coupon bond prices** are contained in the forward rate informations, as bond prices can be written down by integrating over the forward rate between t and T in terms of the risk-neutral process

$$V(t, T) = \exp \left(- \int_t^T f(t, s) ds \right). \tag{17}$$

As we mentioned already, a given model in the HJM model with jump will result in a particular behavior for the short term interest rate. We introduce relation between the short rate process and the forward rate process as follows. In this study, we jump-diffusion version of Hull and White model to reflect this restriction condition. We know the following model for the interest rate r ;

$$dr(t) = a(t)[\theta(t)/a(t) - r(t)]dt + \sigma_r(t)r(t)^\beta dW^Q(t) + Jd\pi(t), \tag{18}$$

where, $\theta(t)$ is a time-dependent drift; $\sigma_r(t)$ is the volatility factor; $a(t)$ is the reversion rate; $dW(t)$ is standard Wiener process; $d\pi(t)$ represents the Poisson process.

Theorem 4. *Let be the jump-diffusion process in short rate $r(t)$ is the equation (18). Let be the volatility form is*

$$\sigma_f(t, T) = \sigma_r(t)(\sqrt{r(t)})^\beta \eta(t, T) \tag{19}$$

with $\eta(t, T) = \exp \left(- \int_t^T a(s) ds \right)$ is deterministic functions. We know the jump-diffusion process in short rate model and the "corresponding" compatible HJM model with jump

$$df(t, T) = \mu_f(t, T)dt + \sigma_f(t, T)dW^Q(t) + Jd\pi(t) \tag{20}$$

where $\mu_f(t, T) = \sigma_f(t, T) \int_t^T \sigma_f(t, s) ds$. Then we obtain the equivalent model is

$$f(0, T) = r(0)\eta(0, T) + \int_0^T \theta(t)\eta(s, T)ds - \int_0^T \sigma_r^2(s)(r(s)^2)^\beta \eta(s, T) \int_s^T (\eta(s, u)du)ds \tag{21}$$

that is, all forward rates are normally distributed. Note that we know that $\beta = 0$ case is an extension of Vasicek's jump diffusion model; the $\beta = 0.5$ case is an extension of CIR jump diffusion model.

Note that the forward rates are normally distributed, which means that the bond prices are log-normally distributed. Both the short term rate and the forward rates can become negative. As above, we obtain the bond price from the theorem 1. By the theorem 2, we drive the relation between the short rate and forward rate.

Corollary 1. *Let be the HJM model with jump of the term structure of interest rate is the stochastic differential equation for forward rate $f(t, T)$ is given by*

$$df(t, T) = \sigma_f(t, T) \int_t^T \sigma_f(t, s) ds dt + \sigma_f(t, T) dW^Q(t) + Jd\pi(t) \tag{22}$$

where, dW_i^Q is the Wiener process generated by an equivalent martingale measure Q and $\sigma_f(t, T) = \sigma_r(t)(\sqrt{r(t)})^\beta \exp\left(-\int_t^T a(s) ds\right)$.

Then the discount bond price $V(t, T)$ for the forward rate is given by the formula

$$V(t, T) = \frac{V(0, T)}{V(0, t)} \exp\left\{-\frac{1}{2} \left(\frac{\int_t^T \sigma_f(t, s) ds}{\sigma_f(t, T)}\right)^2 \int_0^t \sigma_f^2(s, t) ds - \frac{\int_t^T \sigma_f(t, s) ds}{\sigma_f(t, T)} [f(0, t) - r(t)]\right\}$$

with the equation (21).

Note that we know that $\beta = 0$ case is an extension of Vasicek’s jump diffusion model; the $\beta = 0.5$ case is an extension of CIR jump diffusion model.

4 Monte Carlo Simulation of the Term Structure Models with Jump

By and application of Girsanov’s theorem the dependence on the market price of risk can be absorbed into an equivalent martingale measure. Let $W(t), 0 \leq t \leq T$, be a Wiener process on a probability space (Ω, F, P) . Let $\lambda(t), 0 \leq t \leq T$, be a process adapted to this filtration. The Wiener processes $dW^Q(t)$ under the equivalent martingale measure Q are given by $W^Q(t) = W(t) + \int_0^t \lambda(s) ds$ so that

$$dW_i^Q(t) = dW_i(t) + \lambda_i(t) ds.$$

A **risk-neutral measure** Q is any probability measure, equivalent to the market measure P , which makes all discounted bond prices martingales.

We now move on to discuss Monte Carlo simulation. A Monte Carlo simulation of a stochastic process is a procedure for sampling random outcomes for the process. This uses the risk-neutral valuation result. The bond price can be expressed as:

$$V(r_t, t, T) = E_t^Q \left[e^{-\int_t^T r_s ds} | r(t) \right] \quad \text{or} \quad V(f_t, t, T) = E_t^Q \left[e^{-\int_t^T f(t,s) ds} \right] \tag{23}$$

where E_t^Q is the expectations operator with respect to the equivalent risk-neutral measure. Under the equivalent risk-neutral measure, the local expectation hypothesis holds (that is, $E_t^Q \left[\frac{dV}{V} \right]$). To execute the Monte Carlo simulation, we discretize the equations (5) and (12). we divide the time interval $[t, T]$ into m equal time steps of length Δt each. For small time steps, we are entitled to use the discretized version of the risk-adjusted stochastic differential equations (5), (11), and (22):

$$r_j = r_{j-1} + [(\theta \cdot t) - (a \cdot t)r_{j-1} \cdot t - (\lambda \cdot t)(\sigma \cdot t)]\Delta t + (\sigma \cdot t)\varepsilon_j\sqrt{\Delta t} + J_jN_{\Delta t}, \tag{24}$$

$$r_j = r_{j-1} + [(\theta \cdot t) - (a \cdot t)r_{j-1} - (\lambda \cdot t)(\sigma \cdot t)\sqrt{r_{j-1} \cdot t}]\Delta t + (\sigma \cdot t)\sqrt{r_{j-1} \cdot t} \varepsilon_j\sqrt{\Delta t} + J_jN_{\Delta t} \tag{25}$$

and

$$f_j = f_{j-1} + \left[\sigma_f(t, T) \int_t^T \sigma_f(t, s) ds dt \right] \Delta t + \sigma_f(t, T)\varepsilon_j\sqrt{\Delta t} + J_jN_{\Delta t} \tag{26}$$

where $\sigma_f(t, T) = \sigma_r(t)(\sqrt{r(t)})^\beta \exp\left(-\int_t^T a(s)ds\right)$, $j = 1, 2, \dots, m$, ε_j is standard normal variable with $\varepsilon_j \sim N(0, 1)$, and $N_{\Delta t}$ is a Poisson random variable with parameter $h\Delta t$. We know that $\beta = 0$ case is an extension of Vasicek’s jump diffusion model; the $\beta = 0.5$ case is an extension of CIR jump diffusion model. We can investigate the value of the bond by sampling n spot rate paths under the discrete process approximation of the risk-adjusted processes of the equations (24), (25), and (26). The bond price estimate is given by:

$$V(r_t, t, T) = \frac{1}{n} \sum_{i=1}^n \exp\left(-\sum_{j=0}^{m-1} r_{ij}\Delta t\right) \text{ or } V(f_t, t, T) = \frac{1}{n} \sum_{i=1}^n \exp\left(-\sum_{j=0}^{m-1} f_{ij}\Delta t\right),$$

where r_{ij} is the value of the short rate and f_{ij} is the value of the forward rate under the discrete risk-adjusted process within sample path i at time $t + \Delta t$. Numerical methods that are known as Monte Carlo methods can be loosely described as statistical simulation methods, where statistical simulation is defined in quite general terms to be any method that utilizes sequences of random numbers to perform the simulation. The Monte Carlo simulation is clearly less efficient computationally than the numerical method. One measure of the sharpness of the point estimate of the mean is MSE, defined as

$$MSE = \nu/\sqrt{n} \tag{27}$$

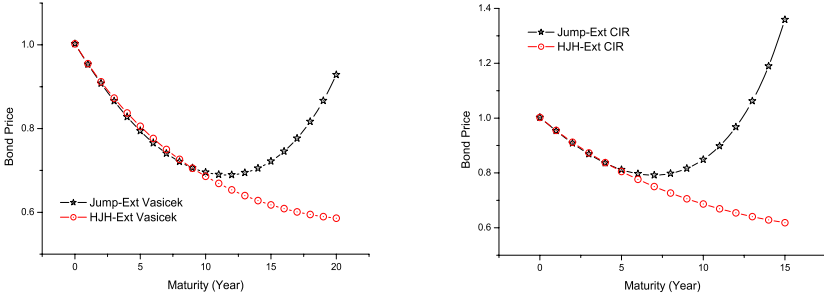
where, ν^2 is the estimate of the variance of bond prices as obtained from n sample paths of the short rate:

$$\nu^2 = \frac{\sum_{i=1}^n \left[\exp\left(-\sum_{j=0}^{m-1} f_{ij}\Delta t\right) - \nu \right]^2}{n - 1}. \tag{28}$$

This reduces the MSE by increasing the value of n . However, highly precise estimates with the brute force method can take a long time to achieve. For the purpose of simulation, we conduct three runs of 1,000 trials each and divide the year into 365 time steps.

5 Experiments

In this section, we investigate the jump-diffusion version of extended Vasicek and CIR model and HJM model with jump. Experiments are consist of the



(a) Bond prices based on extended Vasicek model (b) Bond prices based on extended CIR model

Fig. 1. The various bond prices for the term structure models with jump

numerical method and Monte Carlo simulation. Experiment 1, 2 plot estimated term structure using the various models. In experiment 1, 2, the parameter values are assumed to be $r = 0.05$, $a = 0.5$, $b = 0.05$, $\theta = 0.025$, $\sigma = 0.08$, $\sigma_r = 0.08$, $\lambda = -0.5$, $\gamma = 0.01$, $\mu = 0$, $h = 10$, $t = 0.05$, and $T = 20$. Experiment 3, 4 examine bond prices by the Monte Carlo simulation. In experiment 3, 4, the parameter values are assumed to be $r = 0.05$, $f[0, t] = 0.049875878$, $\sigma_r = 0.08$, $a = 0.5$, $b = 0.05$, $\theta = 0.025$, $\sigma = 0.08$, $\lambda = -0.5$, $\Delta t = (T - t)/m$, $m = 365$, $n = 1000$, $\gamma = 0.01$, $\mu = 0$, $h = 10$, $t = 0.05$, and $T = 20$.

	J-Vasicek	J-E_Vasicek	HJM-E_Vasicek	J-HJM-E_Vasicek
CFS	0.93596	0.953704	0.954902	0.954902
MCS	0.933911	0.95031	0.951451	0.951722
Diff(CFS-MCS)	0.00150833	0.0002868	5.03495E-06	0.000319
Variance	0.00122814	0.00053554	7.09574E-05	0.00178619
MSE	0.000933	0.00286	0.00205	0.003394

Experiment 3. Bond prices estimated by the Monte Carlo simulation for the jump-diffusion and HJM model with jump based on Vasicek model.

	J-CIR	J-E_CIR	HJM-E_CIR	J-HJM-E_CIR
CFS	0.942005	0.953478	0.95491	0.95491
MCS	0.947482	0.951688	0.951456	0.950456
Diff(CFS-MCS)	0.0002863	0.00030634	1.2766E-06	0.0002897
Variance	0.000535	0.0005535	0.000113	0.001702
MSE	-0.005478	0.00179042	0.00345374	0.00444414

Experiment 4. Bond prices estimated by the Monte Carlo simulation for the jump-diffusion and HJM model with jump based on CIR model.

6 Conclusion

Even though Monte Carlo simulation is both harder and conceptually more difficult to implement than the other numerical methods, interest in use of Monte Carlo simulation for bond pricing is getting stronger because of its flexibility in evaluating and handling complicated financial instruments. However, it takes a long time to achieve highly precise estimates with the brute force method. In this paper we investigate bond pricing models and their Monte Carlo simulations with several scenarios. The bond price is humped in the jump versions of the extended Vasicek and CIR models while the bond prices are decreasing functions of the maturity in HJM models with jump.

References

1. C. Ahn and H. Thompson, "Jump-Diffusion Processes and the Term Structure of Interest Rates," *Journal of Finance*, vol. 43, pp. 155-174, 1998.
2. J. Baz and S. R. Das, "Analytical Approximations of the Term Structure for Jump-Diffusion Processes : A Numerical Analysis," *Journal of Fixed Income*, vol. 6(1), pp. 78-86, 1996.
3. D. Beaglehole and M. Tenney, "Corrections and Additions to 'A Nonlinear Equilibrium Model of the Term Structure of Interest Rates'," *Journal of Financial Economics*, vol. 32, pp. 345-353, 1992.
4. E. Briys, "Options, Futures and Exotic Derivatives," John Wiley, 1985.
5. J. C. Cox, J. Ingersoll, and S. Ross, "A Theory of the Term Structure of Interest Rate," *Econometrica*, vol. 53, pp. 385-407, 1985.
6. D. Heath, R. Jarrow, and A. Morton, "Bond Pricing and the Term Structure of Interest Rates," *Econometrica*, vol. 60. no. 1, pp. 77-105, 1992.
7. T. S. Ho and S. Lee, "Term Structure Movements and Pricing Interest Rate Contingent Claims," *Journal of Finance*, vol. 41, pp. 1011-1028, 1986.
8. F. Jamshidian, "An Exact Bond Option Formula," *Journal of Finance*, vol. 44, 1989.
9. J. Frank, and CFA. Fabozzi, "Bond markets Analysis and strategies," Fourth Edition, 2000.
10. J. Hull and A. White, "Pricing Interest Rate Derivative Securities," *Review of Financial Studies*, vol. 3, pp. 573-92, 1990.
11. J. Hull and A. White, "Options, Futures, and Derivatives," Fourth Edition, 2000.
12. F. A. Longstaff and S. Schwartz, "Interest Rate Volatility and the Term structure: A Two-Factor General Equilibrium Model," *Journal of Finance*, vol. 47, pp. 1259-1282, 1992.
13. M. J. Brennan and E. S. Schwartz, "A Continuous Time Approach to the Pricing of Bonds," *Journal of Banking and Finance*, vol. 3, pp. 133-155, 1979.
14. J. Strikwerda, "Finite Difference Schemes and Partial Differential Equations," Wadsworth and Brooks/Cole Advanced Books and Software, 1989.
15. J. W. Drosen, "Pure jump shock models in reliability," *Adv. Appl. Probab.* vol. 18, pp. 423-440, 1986.
16. O. A. Vasicek, "An Equilibrium Characterization of the Term Structure," *Journal of Financial Economics*, vol. 5, pp. 177-188, 1977.

The Ellipsoidal l_p Norm Obnoxious Facility Location Problem

Yu Xia*

The Institute of Statistical Mathematics,
4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan
yuxia@ism.ac.jp

Abstract. We consider locating an obnoxious facility. We use the weighted ellipsoidal l_p norm to accurately measure the distance. We derive the necessary and sufficient conditions for the local optimality and transform the optimality conditions into a system of nonlinear equations. We use Newton's method with perturbed nonmonotone line search to solve the equations. Some numerical experiments are presented.

1 Introduction

We consider the sitting of an obnoxious facility.

Some facilities, such as chemical plants, power plants, air ports, train stations, nuclear reactors, waste dumps, emanate chemical or nuclear pollutants, heat, noise, or magnetic waves. The residents in the region where the obnoxious facility is to be located desire that this facility to be sit as far away as possible from them. The obnoxiousness of the facility is usually an inverse factor of the distance to it. In this paper, we use the maximum weighted sum of distances as the objective. The weights carry the relative importance of existing residence sites.

The dispersion of pollutants is usually affected by meteorology, ground morphology. For instance, hills, tall buildings, rivers, and greens may affect the velocities of winds in different directions in different areas. These effects direct the pollutants to travel in certain paths. As well, forests and mountains may damp the pollution effect. In a series of papers ([3, 4], etc.), it is argued with empirical study that l_p distances weighted by an inflation factor tailored to given regions can better describe the irregularity in the transportation networks such as hills, bends, and is therefore superior to the weighted rectangular and Euclidean norms. This is the same case for pollution distance measure. The impact of the trees, hills, buildings between the pollution source and the residents need to be considered. The curvature of the pollution spread path caused by some geometrical structures can be better described through the rotation of the axes of the coordinates for the residence locations and a proper choice of p . The damping

* Research supported by a postdoctoral fellowship for foreign researchers from JSPS (Japan Society for the Promotion of Science). I thank suggestions and comments of anonymous referees.

factors, such as trees, mountains, can be modeled by the scaling of corresponding axes of the coordinates.

Denote the l_p norm (Minkowski distance of order $p, p > 1$) of a d -dimensional Euclidean space \mathbb{R}^d as

$$l_p(\mathbf{z}) \stackrel{\text{def}}{=} \left[\sum_{i=1}^d |z_i|^p \right]^{1/p} .$$

In this paper, we measure the distances between sites by an ellipsoidal l_p norm distance:

$$l_{pM}(\mathbf{z}) \stackrel{\text{def}}{=} l_p(M\mathbf{z}) ,$$

where M is a linear transformation. It is not hard to see that l_{pM} is a norm when M is nonsingular. Note that the l_{pM} distance is the l_p norm when M is the identity. And it includes Euclidean distance (l_2 -norm distance), Manhattan distance (l_1 norm distance), Chebyshev distance (l_∞ norm distance). It also includes the l_{pb} norm distance (see, for instance [2]):

$$l_{pb}(\mathbf{z}) \stackrel{\text{def}}{=} \left[\sum_{i=1}^n b_i |x_i|^p \right]^{1/p} , \quad b_i > 0 (i = 1, \dots, n) .$$

Obviously the l_{pM} distance measure can better describe the actual transportation networks than the weighted l_p distance or l_{pb} distance can. The p and M may be different for different existing facilities.

Let vectors $\mathbf{f}_1, \dots, \mathbf{f}_n$ represent the n existing residence sites. Let vector \mathbf{x} denote the site where the new facility to be located.

The distances from the obnoxious facility to some essential residence sites $\mathbf{f}_j (j = 1, \dots, s)$ such as hospitals, schools, tourism spots, should be above some thresholds r_j , due to some legal or environmental considerations. Some of the essential residence sites $\mathbf{f}_j (j = 1, \dots, s)$ may or may not be the same as the existing sites $\mathbf{f}_i (i = 1, \dots, n)$ included in the maxisum objective. Let w_i denote the weight on location $\mathbf{f}_i (i = 1, \dots, n)$, which is associated with the number and importance of residents there. Of some type of residents, such as patients, children, the weights might be higher than those of others.

The model we are considering is the following.

$$\max_{\mathbf{x}} \quad \sum_{i=1}^n w_i \|M_i(\mathbf{x} - \mathbf{f}_i)\|_{p_i} \tag{1a}$$

$$\text{s.t.} \quad \mathbf{Ax} \leq \mathbf{b} \tag{1b}$$

$$\|M_{n+j}(\mathbf{x} - \mathbf{f}_{n+j})\|_{p_{n+j}} \geq r_j \quad (j = 1, \dots, s) \tag{1c}$$

The objective is convex. Therefore, (1) is not easy to solve, since it is a maximization problem. In addition, the constraints are not convex. Next, we consider its dual and optimality conditions. Our aim is to reformulate the optimality conditions into a system of equations which can then be solved by Newton's method.

The remaining of the paper is organized as follows. In §2, we derive the optimality conditions for (1) and then reformulate the optimality conditions into

a system of equations. In §3, we present our Newton’s method with perturbed nonmonotone line search algorithm. In §4, we give some numerical examples of our algorithm.

2 The Optimality Conditions

In this part, we consider the Lagrangian dual of (1) and derive its optimality conditions.

For each $p_i \geq 0$, we define a scalar p_i satisfying

$$\frac{1}{p_i} + \frac{1}{q_i} = 1 .$$

By Hölder’s inequality,

$$\max_{\mathbf{x}} w_i \|M_i(\mathbf{x} - \mathbf{f}_i)\|_{p_i} = \min_{(\|\mathbf{z}_i\|_{q_i} \geq w_i)} \max_{\mathbf{x}} \mathbf{z}_i^T M_i(\mathbf{x} - \mathbf{f}_i) ,$$

with $\|\mathbf{z}_i\|_{q_i} = w_i$, $\|M_i(\mathbf{x} - \mathbf{f}_i)\|_{p_i}^{p_i} |z_{il}|^{q_i} = w_i^{q_i} |M_{il}(\mathbf{x} - \mathbf{f}_i)|^{p_i}$, $\text{sign}(\mathbf{z}_{il}) = \text{sign}[M_{il}(\mathbf{x} - \mathbf{f}_i)]$ when $M_i(\mathbf{x} - \mathbf{f}_i) \neq \mathbf{0}$; and $\|\mathbf{z}_i\|_{q_i} \geq w_i$ when $M_i(\mathbf{x} - \mathbf{f}_i) = \mathbf{0}$.

Hence, we have that the dual to (1) is the following.

$$\begin{aligned} \min_{\lambda} \min_{\mathbf{0}, \boldsymbol{\eta}} \max_{\mathbf{x}} \sum_{i=1}^n w_i \|M_i(\mathbf{x} - \mathbf{f}_i)\|_{p_i} - \lambda^T (A\mathbf{x} - \mathbf{b}) + \sum_{j=1}^s \eta_j \left(\|M_{n+j}(\mathbf{x} - \mathbf{f}_{n+j})\|_{p_{n+j}} \right. \\ \left. - r_j \right) = \lambda \min_{\mathbf{0}, \boldsymbol{\eta}} \min_{\substack{\mathbf{z}_i \quad q_i \quad w_i \\ (i=1, \dots, n) \\ \mathbf{z}_j \quad q_j \quad \eta_j \\ (j=n+1, \dots, n+s)}} \max_{\mathbf{x}} \left[\sum_{i=1}^n \mathbf{z}_i^T M_i(\mathbf{x} - \mathbf{f}_i) - \lambda^T (A\mathbf{x} - \mathbf{b}) \right. \\ \left. + \sum_{j=1}^s \mathbf{z}_{j+n}^T M_{j+n}(\mathbf{x} - \mathbf{f}_{j+n}) - \sum_{j=1}^s \eta_j r_j \right] = \lambda \min_{\mathbf{0}, \boldsymbol{\eta}} \min_{\substack{\mathbf{z}_i \quad q_i \quad w_i \\ (i=1, \dots, n) \\ \mathbf{z}_j \quad q_j \quad \eta_j \\ (j=n+1, \dots, n+s)}} \max_{\mathbf{x}} \\ \left[\left(\sum_{i=1}^{n+s} \mathbf{z}_i^T M_i - \lambda^T A \right) \mathbf{x} - \sum_{i=1}^{n+s} \mathbf{z}_i^T M_i \mathbf{f}_i + \lambda^T \mathbf{b} - \sum_{j=1}^s \eta_j r_j \right] . \end{aligned}$$

In the above expression, $\sum_{i=1}^{n+s} \mathbf{z}_i^T M_i = \lambda^T A$; otherwise, the inner maximization would be unbounded.

Therefore, the dual to (1) is

$$\begin{aligned} \min_{\lambda, \boldsymbol{\eta}, \mathbf{z}_i} & - \sum_{i=1}^{n+s} \mathbf{z}_i^T M_i \mathbf{f}_i + \lambda^T \mathbf{b} - \sum_{j=1}^s \eta_j r_j \\ \text{s.t.} & \sum_{i=1}^{n+s} M_i^T \mathbf{z}_i - A^T \lambda = \mathbf{0} \\ & \|\mathbf{z}_i\|_{q_i} \geq w_i \quad (i = 1, \dots, n) \\ & \|\mathbf{z}_j\|_{q_j} \geq \eta_j \quad (j = n + 1, \dots, n + s) \\ & \lambda \geq \mathbf{0} \\ & \boldsymbol{\eta} \geq \mathbf{0} \end{aligned} \tag{2}$$

In addition, we have $\|\mathbf{z}_i\|_{q_i} = w_i$, $\|M_i(\mathbf{x} - \mathbf{f}_i)\|_{p_i}^{p_i} |z_{il}|^{q_i} = w_i^{q_i} |M_{il}(\mathbf{x} - \mathbf{f}_i)|^{p_i}$, $\text{sign}(\mathbf{z}_{il}) = \text{sign}[M_{il}(\mathbf{x} - \mathbf{f}_i)]$ when $M_i(\mathbf{x} - \mathbf{f}_i) \neq \mathbf{0}$ for $i = 1, \dots, n$. And $\|\mathbf{z}_j\|_{q_j} =$

$\eta_j, \|M_j(\mathbf{x} - \mathbf{f}_j)\|_{p_j}^{p_j} |z_{jl}|^{q_j} = \eta_j^{q_j} |M_{jl}(\mathbf{x} - \mathbf{f}_j)|^{p_j}, \text{sign}(z_{jl}) = \text{sign}[M_{jl}(\mathbf{x} - \mathbf{f}_j)]$ when $M_i(\mathbf{x} - \mathbf{f}_j) \neq \mathbf{0}$ for $j = 1, \dots, s$. We also have $\lambda_i(A_i\mathbf{x} - b_i) = 0 (i = 1, \dots, m), \eta_j (\|M_{n+j}(\mathbf{x} - \mathbf{f}_{n+j})\|_{p_{n+j}} - r_j) = 0, (j = 1, \dots, s)$.

The new facility is required not to be located in existing residence sites. Then the objective is differentiable at optimum. Therefore, the reverse convex constraint qualification is satisfied. Thus, there is no duality gap between (1) and (2) (see [5]).

We conclude that the necessary and sufficient local optimality conditions for (1) are:

$$\begin{aligned} & \sum_{i=1}^{n+s} M_i^T \mathbf{z}_i - A^T \boldsymbol{\lambda} = \mathbf{0} , \\ & \eta_{j-n} \left(r_{j-n}^{p_j} - \|M_j(\mathbf{x} - \mathbf{f}_j)\|_{p_j}^{p_j} \right) = 0 \quad (j = n + 1, \dots, n + s) , \\ & \eta_j \geq 0 \quad (j = 1, \dots, s) , \\ & \|M_{n+j}(\mathbf{x} - \mathbf{f}_{n+j})\|_{p_{n+j}} \geq r_j \quad (j = 1, \dots, s) , \\ & \lambda_i(b_i - A_i\mathbf{x}) = 0 \quad (i = 1, \dots, m) , \\ & \lambda_i \geq 0 \quad (i = 1, \dots, m) , \\ & A_i\mathbf{x} \leq b_i \quad (i = 1, \dots, m) , \\ & \alpha_i |z_{il}|^{q_i} = |M_{il}(\mathbf{x} - \mathbf{f}_i)|^{p_i} \quad (i = 1, \dots, n; l = 1, \dots, d_i) , \\ & \alpha_j |z_{jl}|^{q_j} = \eta_{j-n} |M_{jl}(\mathbf{x} - \mathbf{f}_j)|^{p_j} \quad (j = 1 + n, \dots, n + s; l = 1, \dots, d_j) , \\ & \text{sign}(z_{il}) = \text{sign}[M_{il}(\mathbf{x} - \mathbf{f}_i)] \quad (i = 1, \dots, n + s; l = 1, \dots, d_i) , \\ & \alpha_i(w_i - \|\mathbf{z}_{il}\|_{p_i}) = 0 \quad (i = 1, \dots, n) , \\ & \|\mathbf{z}_i\|_{q_i} \geq w_i \quad (i = 1, \dots, n) , \\ & \alpha_{n+j}(\eta_j - \|\mathbf{z}_{n+j}\|_{p_{n+j}}) = 0 \quad (j = 1, \dots, s) , \\ & \|\mathbf{z}_{n+j}\|_{q_{n+j}} \geq \eta_j \quad (j = 1, \dots, s) . \end{aligned}$$

We define $\text{sign}(0) = \text{sign}(a)$ to be true for all $a \in \mathbb{R}$.

Note that the above system is not easy to solve, because it includes some inequalities.

We use some nonlinear complementarity functions to reformulate the complementarity. Specially, we use the min function, since it is the simplest. Observe that $\min(a, b) = 0$ iff $a, b \geq 0$ and at least one of a and b is 0.

We assume that $\mathbf{f}_i (i = 1, \dots, n + s)$ doesn't satisfy $A\mathbf{x} \leq \mathbf{b}$. This means that the region where the obnoxious facility to be sit doesn't include existing residence sites.

In the formulation, we also distinguish between $p_i \geq 2$ and $p_i < 2$ to avoid some nondifferentiable points.

From $\frac{1}{p_i} + \frac{1}{q_i} = 1, (i = 1, \dots, n + s)$, we have

$$\begin{aligned} & p_i \geq 2 \Rightarrow q_i \leq 2, \quad p_i \leq 2 \Rightarrow q_i \geq 2 ; \\ & \frac{p_i}{q_i} = p_i - 1 = \frac{1}{q_i - 1}, \quad \frac{q_i}{p_i} = q_i - 1 = \frac{1}{p_i - 1} . \end{aligned}$$

We then transform the optimality conditions into the following system of equations.

$$\sum_{i=1}^{n+s} M_i^T \mathbf{z}_i - A^T \boldsymbol{\lambda} = \mathbf{0} \tag{3a}$$

$$\min \left[\eta_{j-n}, \|M_j(\mathbf{x} - \mathbf{f}_j)\|_{p_j}^{p_j} - r_{j-n}^{p_j} \right] = 0 \quad (j = n + 1, \dots, n + s), \tag{3b}$$

$$\min (\lambda_i, b_i - A_i \mathbf{x}) = 0 \quad (i = 1, \dots, m), \tag{3c}$$

$$\left(\sum_{l=1}^{d_i} |M_{il}(\mathbf{x} - \mathbf{f}_i)|^{p_i} \right)^{\frac{1}{q_i}} z_{il} - w_i |M_{il}(\mathbf{x} - \mathbf{f}_i)|^{\frac{p_i}{q_i}} \text{sign}(M_{il}(\mathbf{x} - \mathbf{f}_i)) = 0 \tag{3d}$$

$(i \in \{1, \dots, n\}, p_i \geq 2; l = 1, \dots, d_i)$

$$\left(\sum_{l=1}^{d_i} |M_{il}(\mathbf{x} - \mathbf{f}_i)|^{p_i} \right)^{\frac{1}{p_i}} |z_{il}|^{\frac{q_i}{p_i}} \text{sign}(z_{il}) - w_i^{\frac{q_i}{p_i}} M_{il}(\mathbf{x} - \mathbf{f}_i) = 0 \tag{3e}$$

$(i \in \{1, \dots, n\}, p_i < 2; l = 1, \dots, d_i)$

$$\left(\sum_{l=1}^{d_j} |M_{jl}(\mathbf{x} - \mathbf{f}_j)|^{p_j} \right)^{\frac{1}{q_j}} z_{jl} - \eta_{j-n} |M_{jl}(\mathbf{x} - \mathbf{f}_j)|^{\frac{p_j}{q_j}} \text{sign}(M_{jl}(\mathbf{x} - \mathbf{f}_j)) = 0 \tag{3f}$$

$(j \in \{n + 1, \dots, n + s\}, p_j \geq 2; l = 1, \dots, d_j)$

$$\left(\sum_{l=1}^{d_j} |M_{jl}(\mathbf{x} - \mathbf{f}_j)|^{p_j} \right)^{\frac{1}{p_j}} |z_{jl}|^{\frac{q_j}{p_j}} \text{sign}(z_{jl}) - \eta_{j-n}^{\frac{q_j}{p_j}} M_{jl}(\mathbf{x} - \mathbf{f}_j) = 0 \tag{3g}$$

$(j \in \{n + 1, \dots, n + s\}, p_j < 2; l = 1, \dots, d_j)$

3 The Algorithm

Let F represent the left-hand-side of (3). Let $\Psi \stackrel{\text{def}}{=} \frac{F \cdot F}{2}$. Any global optimization method that locates a global minimal solution to Ψ finds a local solution to (1). Unfortunately, Ψ is not differentiable everywhere. We then use the gradient decent method by perturbed nonmonotone line search [6] to skip the nonsmooth points to find a global minimum of Ψ .

Denote $\mathbf{u} \stackrel{\text{def}}{=} (\mathbf{x}; \boldsymbol{\lambda}; \boldsymbol{\eta}; \mathbf{z})$. Let $\Delta \mathbf{u}$ represent the Newton's direction to (3). Below is the algorithm.

The Algorithm

Initialization. Set constants $s > 0, 0 < \sigma < 1, \beta \in (0, 1), \gamma \in (\beta, 1), nml \geq 1$.

For each $k \geq 0$, assume Ψ is differentiable at \mathbf{u}^k . Set $k = 0$.

Do while. $\|F\|_\infty \geq \text{opt}, \|\mathbf{u}^{k+1} - \mathbf{u}^k\|_\infty \geq \text{septol}$, and $k \leq \text{itlimit}$.

1. Find the Newton's direction for (3): $\Delta \mathbf{u}^k$.
2. (a) Set $\alpha^{k,0} = s, i = 0$.

(b) Find the smallest nonnegative integer l for which

$$\Psi(\mathbf{u}^k) - \Psi(\mathbf{u}^k + \beta^l \alpha^{k,i} \Delta \mathbf{u}^k) \geq_{0 \leq j \leq m(k)} -\sigma \beta^l \alpha^{k,i} \nabla \Psi(\mathbf{u}^j)^T \Delta \mathbf{u}^k.$$

where $m(0) = 0$ and $0 \leq m(k) \leq \min[m(k-1) + 1, nml]$.

(c) If Ψ is nondifferentiable at $(\mathbf{u}^k + \beta^l \alpha^{k,i} \Delta \mathbf{u}^k)$, find $t \in [\gamma, 1)$ so that Ψ is differentiable at $(\mathbf{u}^k + t\beta^l \alpha^{k,i} \Delta \mathbf{u}^k)$, set $\alpha^{k,i+1} = t\beta^l \alpha^{k,i}$, $i + 1 \rightarrow i$, go to step 2b.

Otherwise, set $\alpha^k = \beta^l \alpha^{k,i}$, $\mathbf{u}^{k+1} = \mathbf{u}^k + \alpha^k \Delta \mathbf{u}$, $k + 1 \rightarrow k$.

4 Numerical Experiments

We adopt the suggested parameters in [1]. The machine accuracy of the computer running the code is $\epsilon = 2.2204e - 16$. Our computer program stops either $\|F\|_\infty < opt = \epsilon^{1/3} = 6.0555e - 5$, or the infinity norm of the Newton's direction is less than $steptol = \epsilon^{2/3}$; or the number of iterations exceeds $itlimit = 100$. We set $s = 1$, $\beta = \frac{1}{2}$, $\sigma = 1.0e - 4$, $nml = 10$.

Below is an example with 10 existing facilities and 2 essential facilities from which the obnoxious facility must be away for some minimal distance.

The 10 existing facilities are:

$$\begin{aligned} f_1 &= (0.8351697, 0.9708701, 0.4337257), f_2 = (0.5029927, 0.4754272, 0.7399495), \\ f_3 &= (0.9887191, 0.4000630, 0.6804281), f_4 = (0.9093165, 0.5075840, 0.8947370), \\ f_5 &= (0.4425517, 0.3319756, 0.0674839), f_6 = (0.5963061, 0.1664579, 0.1948914), \\ f_7 &= (0.7186201, 0.7451580, 0.5479852), f_8 = (0.4763083, 0.9978569, 0.5943900), \\ f_9 &= (0.4766319, 0.0927731, 0.4870974), f_{10} = (0.1512887, 0.2611954, 0.6834821). \end{aligned}$$

The 2 essential facilities are

$$f_{11} = (0.5290907, 0.1980252, 0.1012210), f_{12} = (0.5800074, 0.8072991, 0.1645741).$$

The weights in the objective are

$$\mathbf{w} = (0.1093683, 0.2339055, 0.5295364, 0.2302270, 0.4429267, 0.3831922, 0.1667756, 0.7351673, 0.1278835, 0.4373618).$$

The p for the ellipsoidal norms are

$$p = (2.1, 2.3, 1.5, 3.1, 1.9, 1.9, 1.9, 1.9, 1.9, 1.7, 3.3, 1.9).$$

The linear transformation matrices for the ellipsoidal norms are:

$$\begin{aligned} M_1 &= \begin{bmatrix} 0.3692412 & 0.7397360 & 0.3671468 \\ 0.3451773 & 0.5917062 & 0.7667686 \\ 0.7153005 & 0.3233307 & 0.5691886 \end{bmatrix} & M_2 &= \begin{bmatrix} 0.5969909 & 0.1652245 & 0.9320327 \\ 0.1077514 & 0.0023804 & 0.9944936 \\ 0.9073230 & 0.5631396 & 0.9684730 \end{bmatrix} \\ M_3 &= \begin{bmatrix} 0.6143491 & 0.1460124 & 0.7866041 \\ 0.0613802 & 0.7311101 & 0.6258767 \\ 0.5635103 & 0.9968714 & 0.8857684 \end{bmatrix} & M_4 &= \begin{bmatrix} 0.4259558 & 0.1795297 & 0.3595944 \\ 0.0765066 & 0.1897343 & 0.1458442 \\ 0.4611495 & 0.7927964 & 0.3008546 \end{bmatrix} \end{aligned}$$

$$\begin{aligned}
 M_5 &= \begin{bmatrix} 0.8462288 & 0.3543530 & 0.9533479 \\ 0.8034236 & 0.5255093 & 0.0520969 \\ 0.0296419 & 0.9122241 & 0.2724550 \end{bmatrix} & M_6 &= \begin{bmatrix} 0.6154476 & 0.0826227 & 0.3935258 \\ 0.8348387 & 0.2062091 & 0.8198316 \\ 0.3425691 & 0.6444498 & 0.2284179 \end{bmatrix} \\
 M_7 &= \begin{bmatrix} 0.9841331 & 0.6252488 & 0.3162361 \\ 0.9204588 & 0.0062540 & 0.2861219 \\ 0.8010166 & 0.2107349 & 0.6545077 \end{bmatrix} & M_8 &= \begin{bmatrix} 0.5306548 & 0.3486466 & 0.8227931 \\ 0.7473973 & 0.7774893 & 0.1455634 \\ 0.2803321 & 0.3570228 & 0.5587755 \end{bmatrix} \\
 M_9 &= \begin{bmatrix} 0.1136328 & 0.6820894 & 0.9157958 \\ 0.7598759 & 0.9467710 & 0.9630559 \\ 0.9203295 & 0.4310774 & 0.8603421 \end{bmatrix} & M_{10} &= \begin{bmatrix} 0.1660018 & 0.3996167 & 0.1370780 \\ 0.5331381 & 0.8280759 & 0.8089569 \\ 0.6531517 & 0.2877654 & 0.2869363 \end{bmatrix} \\
 M_{11} &= \begin{bmatrix} 0.8655222 & 0.8695969 & 0.2533491 \\ 0.7964027 & 0.0003685 & 0.8457027 \\ 0.1553411 & 0.6526919 & 0.8278037 \end{bmatrix} & M_{12} &= \begin{bmatrix} 0.8813783 & 0.2787139 & 0.8239571 \\ 0.9325132 & 0.1975635 & 0.4511900 \\ 0.5321329 & 0.7430964 & 0.5640829 \end{bmatrix}.
 \end{aligned}$$

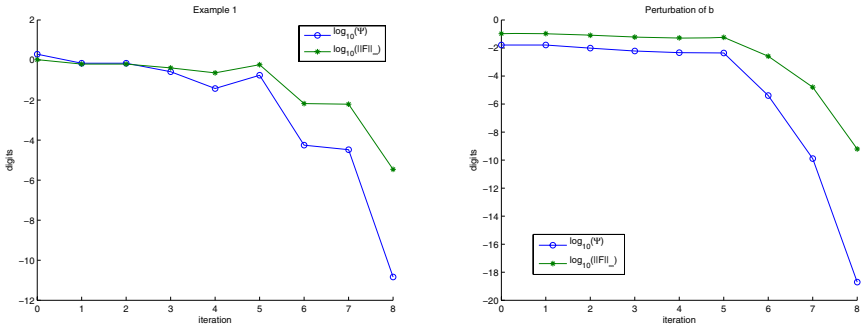
The coefficients for the linear constraints are

$$\begin{aligned}
 A &= \begin{pmatrix} 0.0482753 & 0.4878658 & 0.0619276 \\ -0.8904399 & -0.8477093 & -0.4979966 \\ 0.8629468 & 0.5048775 & 0.7739134 \end{pmatrix}, \\
 \mathbf{b} &= (0.3484993, -1.0367463, 1.0501067)^T.
 \end{aligned}$$

The minimal distances from the essential sites are

$$\mathbf{r} = (0.7952982, 0.1178979)^T.$$

We start from a zero solution: $\mathbf{x} = \mathbf{0}$, $\mathbf{z} = \mathbf{0}$, $\boldsymbol{\lambda} = \mathbf{0}$, $\boldsymbol{\eta} = \mathbf{0}$. The iterates are summarized in the figure “Example 1”. In the figure, x-axis represents the iteration number. The blue plot depicts $\log_{10}(\Psi)$. The green plot depicts $\log_{10}(\|F\|_\infty)$.



Assume some data need to be modified, due to some previous measure error or the availability of more advanced measuring instruments. We then use the old solution as the starting point to solve the new instances by Newton’s method, since Newton’s method has locally Q-quadratic convergence rate.

We perturb \mathbf{b} by some random number in $(-0.5, 0.5)$ to

$$\mathbf{b} = (0.4528631, -0.9323825, 1.1544705)^T.$$

Then we use the perturbed nonmonotone Newton’s method to solve the problem with starting point being the solution to ‘Example 1’. The iterations are summarized in the figure “Perturbation of b”.

We also randomly perturb each element of A in the range $(-0.5, 0.5)$, to

$$A = \begin{pmatrix} -0.0613061 & 0.3782844 & -0.0476538 \\ -1.0000212 & -0.9572907 & -0.6075779 \\ 0.7533654 & 0.3952961 & 0.6643320 \end{pmatrix};$$

perturb each weight w in the range $(-0.5, 0.5)$ to

$$\mathbf{w} = (0.7894956, 0.9140328, 1.2096637, 0.9103542, 1.1230539, 1.0633194, 0.8469029, 1.4152946, 0.8080108, 1.1174891);$$

perturb each existing sites \mathbf{f} randomly in the range $(-0.5, 0.5)$, to

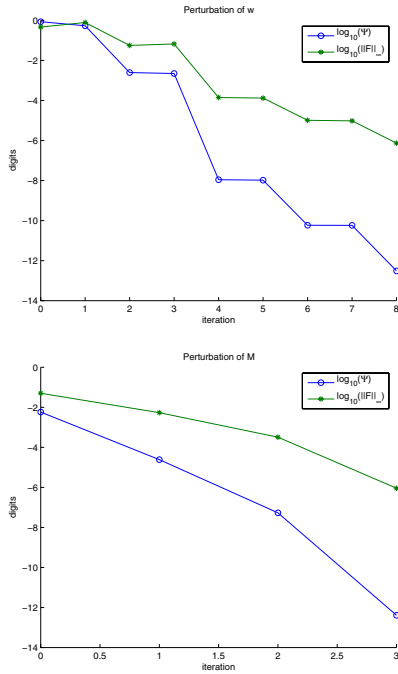
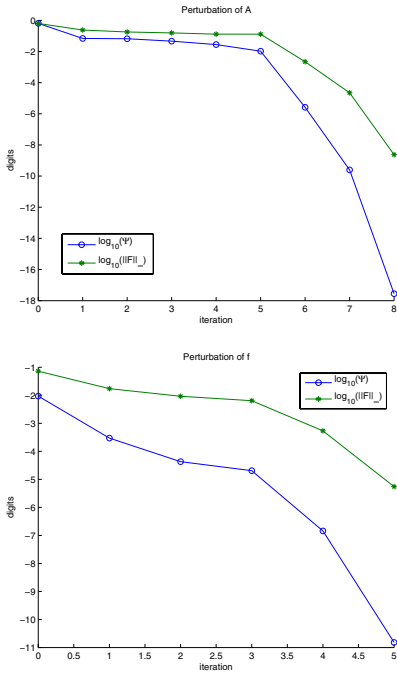
$$\begin{aligned} f_1 &= (0.7483169, 0.8840173, 0.3468729)^T & f_2 &= (0.4161400, 0.3885744, 0.6530967)^T \\ f_3 &= (0.9018664, 0.3132102, 0.5935753)^T & f_4 &= (0.8224637, 0.4207312, 0.8078842)^T \\ f_5 &= (0.3556989, 0.2451229, -0.0193689)^T & f_6 &= (0.5094533, 0.0796051, 0.1080387)^T \\ f_7 &= (0.6317674, 0.6583052, 0.4611324)^T & f_8 &= (0.3894556, 0.9110042, 0.5075373)^T \\ f_9 &= (0.3897792, 0.0059204, 0.4002447)^T & f_{10} &= (0.0644359, 0.1743426, 0.5966293)^T \\ f_{11} &= (0.4422379, 0.1111724, 0.0143683)^T & f_{12} &= (0.4931546, 0.7204463, 0.0777213)^T; \end{aligned}$$

perturb each element of the linear matrices for the ellipsoidal norm randomly by a number in $(-0.5, 0.5)$ to

$$\begin{aligned} M_1 &= \begin{bmatrix} 0.4750002 & 0.8454950 & 0.4729058 \\ 0.4509363 & 0.6974652 & 0.8725276 \\ 0.8210595 & 0.4290897 & 0.6749476 \end{bmatrix} & M_2 &= \begin{bmatrix} 0.7027499 & 0.2709835 & 1.0377917 \\ 0.2135104 & 0.1081394 & 1.1002526 \\ 1.0130820 & 0.6688986 & 1.0742320 \end{bmatrix} \\ M_3 &= \begin{bmatrix} 0.7201081 & 0.2517714 & 0.8923631 \\ 0.1671392 & 0.8368691 & 0.7316357 \\ 0.6692693 & 1.1026304 & 0.9915274 \end{bmatrix} & M_4 &= \begin{bmatrix} 0.5317148 & 0.2852887 & 0.4653534 \\ 0.1822656 & 0.2954933 & 0.2516032 \\ 0.5669084 & 0.8985554 & 0.4066136 \end{bmatrix} \\ M_5 &= \begin{bmatrix} 0.9519878 & 0.4601120 & 1.0591069 \\ 0.9091826 & 0.6312682 & 0.1578559 \\ 0.1354009 & 1.0179831 & 0.3782140 \end{bmatrix} & M_6 &= \begin{bmatrix} 0.7212066 & 0.1883817 & 0.4992848 \\ 0.9405977 & 0.3119681 & 0.9255906 \\ 0.4483281 & 0.7502088 & 0.3341769 \end{bmatrix} \\ M_7 &= \begin{bmatrix} 1.0898921 & 0.7310078 & 0.4219951 \\ 1.0262178 & 0.1120130 & 0.3918809 \\ 0.9067756 & 0.3164939 & 0.7602667 \end{bmatrix} & M_8 &= \begin{bmatrix} 0.6364138 & 0.4544056 & 0.9285521 \\ 0.8531563 & 0.8832483 & 0.2513224 \\ 0.3860911 & 0.4627818 & 0.6645345 \end{bmatrix} \\ M_9 &= \begin{bmatrix} 0.2193918 & 0.7878484 & 1.0215548 \\ 0.8656349 & 1.0525300 & 1.0688149 \\ 1.0260885 & 0.5368364 & 0.9661011 \end{bmatrix} & M_{10} &= \begin{bmatrix} 0.2717608 & 0.5053757 & 0.2428370 \\ 0.6388970 & 0.9338349 & 0.9147159 \\ 0.7589107 & 0.3935244 & 0.3926953 \end{bmatrix} \\ M_{11} &= \begin{bmatrix} 0.9712812 & 0.9753558 & 0.3591081 \\ 0.9021617 & 0.1061274 & 0.9514617 \\ 0.2611001 & 0.7584509 & 0.9335627 \end{bmatrix} & M_{12} &= \begin{bmatrix} 0.9871373 & 0.3844729 & 0.9297161 \\ 1.0382722 & 0.3033225 & 0.5569490 \\ 0.6378919 & 0.8488554 & 0.6698419 \end{bmatrix}. \end{aligned}$$

These instances are then solved by the perturbed nonmonotone Newton’s method starting from the solution to ‘Example 1’. The iterations are summarized

in figures “Perturbation of A”, “Perturbation of w”, “Perturbation of f”, and “Perturbation of M” respectively.



The above instances show the Q-quadratic convergence rate of the Newton’s method.

To find a global solution to (1), randomly restarting of Newton’s method for (1) can be used.

References

1. John E. Dennis, Jr. and Robert B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. Prentice Hall Series in Computational Mathematics. Prentice Hall Inc., Englewood Cliffs, NJ, 1983.
2. J. Fernández, P. Fernández, and B. Pelegrin. Estimating actual distances by norm functions: a comparison between the $l_{k,p,\theta}$ -norm and the $l_{b_1,b_2,\theta}$ -norm and a study about the selection of the data set. *Comput. Oper. Res.*, 29(6):609–623, 2002.
3. R.F. Love and JG Morris. Modelling inter-city road distances by mathematical functions. *Operational Research Quarterly*, 23:61–71, 1972.
4. R.F. Love and JG Morris. Mathematical models of road travel distances. *Management Science*, 25:130–139, 1979.
5. Olvi L. Mangasarian. *Nonlinear programming*, volume 10 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994.
6. Yu Xia. An algorithm for perturbed second-order cone programs. Technical Report AdvOI-Report No. 2004/17, McMaster University, 2004.

On the Performance of Recovery Rate Modeling

J. Samuel Baixauli¹ and Susana Alvarez²

¹ Department of Management and Finance,
University of Murcia, Spain
sbaixaul@um.es

² Department of Quantitative Methods for the Economy,
University of Murcia, Spain
salvarez@um.es

Abstract. To ensure accurate predictions of loss given default it is necessary to test the goodness-of-fit of the recovery rate data to the Beta distribution, assuming that its parameters are unknown. In the presence of unknown parameters, the Cramer-von Mises test statistic is neither asymptotically distribution free nor parameter free. In this paper, we propose to compute approximated critical values with a parametric bootstrap procedure. Some simulations show that the bootstrap procedure works well in practice.

1 Introduction

The probability distribution function of recovery rates is generally unknown. Hence, a probability distribution function that matches quite well the shape of the underlying distribution function should be chosen. In credit risk studies a Beta distribution is usually assumed in order to make predictions. The shape of the Beta distribution depends on the choice of its two parameters p and q . A nonparametric goodness-of-fit test based on a Cramer-von Mises type test statistic (CVM) can be carried out in order to test if the Beta distribution describes reasonably well the empirical shape of the recovery rates, assuming that p and q are unknown. However, in presence of unknown parameters the asymptotic distribution of CVM is not distribution free and it is not even asymptotically parameter free. Consequently, the standard tables used for the traditional Cramer-von Mises test statistic are not longer valid.

In this paper, we implement the test of the null hypothesis “the distribution of the recovery rates is $B(p,q)$ ” versus the alternative “the distribution of the recovery rates is not of this type” using parametric bootstrap methodology and we design simulation experiments to check for the power of such test.

The paper is organized as follows. In Section 1 we describe briefly the problem of financial modeling in credit risk management. In Section 2 we present the statistical problem of carrying out nonparametric goodness-of-fit tests in presence of a vector of unknown parameters. In Section 3 we describe a parametric bootstrap procedure to implement the nonparametric goodness-of-fit test introduced in Section 2. We check that the proposed bootstrap procedure works reasonably well in practice by means of some Monte Carlo experiments. In Section 4 we conclude.

2 Financial Modeling Problem in Credit Risk Management

Recovery rate (R_i) is the amount that a creditor would receive in final satisfaction of the claims on a defaulted credit. In general, this is a percentage of the debt’s par value, which is the most common practice. Loss given default (LGD) is defined as (1-recovery rate). Accurate LGD estimates are fundamental for provisioning reserves for credit losses, calculating risk capital and determining fair pricing for credit risky obligations.

Estimates of LGD have usually been done by traditional methodologies of historical averages segmented by debt type (loans, bonds, stocks, etc.) and seniority (secured, senior unsecured, subordinate, etc.). To use the historical average methodology has several drawbacks, mainly because this methodology is characterized by using the same estimate of recovery irrespective of the horizon over which default might occur, what implies to ignore the credit cycle. Additionally, historical average methodology is updated infrequently; thus, new data have a small impact on longer-term averages. Recently, an alternative methodology appears with the purpose to predict LGD, the Moody’s KMV model (MKMV) [6] and [7].

MKMV model uses several explanatory factors to predict LGD and it assumes that the dependent variable of the regression model follows a Beta distribution. However, there is no theoretical reason that this is the correct shape of the dependent variable. MKMV model assumes the Beta distribution because it might be a reasonable description of the behavior of the recovery rates since it is one of the few common “named” distributions that give probability 1 to a finite interval, here taken to be (0,1), corresponding to 100% loss or zero loss. Additionally, it has great flexibility in the sense that it is not restricted to being symmetrical. The Beta distribution is indexed by two parameters and its probability density function (pdf) is given by:

$$f(x) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1}, \quad 0 \leq x \leq 1, \quad p > 0, \quad q > 0, \quad (1)$$

where, $B(p,q)$ denotes the beta function, p is the shape parameter and q is the scale parameter. Application of a Beta distribution in LGD models can be found in [5] and [9], between others.

As the parameters p and q vary, the Beta distribution takes on different shapes. In this paper, we fix the parameter values in order to represent two scenarios since different recovery rates across securities have been observed [2]. In the first scenario the probability density functions become more concentrated to the left. In the second scenario, the probability density functions become more concentrated to the right. These scenarios correspond to two types of securities: securities with low recovery rate and securities with high recovery rate, respectively.

In order to regress the recovery rates on the explanatory variables, MKMV model converts the ‘assumed’ Beta distributed recovery values to a more normally distributed dependent variable using the normal quantile transformation, defined as follows:

$$\text{Dependent variable} = \tilde{R}_i = N^{-1}(\text{Beta}(R_i, \hat{p}, \hat{q})) , \quad (2)$$

where, N^{-1} is the inverse of the normal distribution function.

Nevertheless, the assumption of Beta distribution is questionable since other distribution functions can be considered as descriptive models for the distribution of recovery rates. Hence, the usual assumption of Beta distribution could imply the fact of the recovery rates contained measurement error. Conceptually, in the measurement error case, the dependent variable is mismeasured. On consequence, accurate LGD modeling requires specifying the correct distributional model instead of assuming the Beta distribution systematically.

3 Goodness-of-Fit Tests in Presence of Some Unknown Parameters

Let R_1, R_2, \dots, R_n be independent and identically (i.i.d.) random variables and let $F(x, \theta)$ be a continuous distribution function, known except for an s -dimensional parameter vector θ . We are interested in testing the hypotheses, H_0 : The distribution of R_i is $F(\cdot, \theta)$ versus H_1 : The distribution of R_i is not $F(\cdot, \theta)$.

The nonparametric procedure for testing the null hypothesis “the distribution of R_i is $F(\cdot, \theta)$ ” versus the alternative “the distribution of R_i is not of this type” consists in comparing empirical distribution function with theoretical distribution function, substituting the unknown parameter vector θ by an estimate $\hat{\theta}$. In this context, the CVM test statistic is defined as:

$$W_n^2 = \int_{\mathcal{R}} n [F_n(x) - F(x, \hat{\theta})]^2 dF_n(x) = \sum_{i=1}^n \{F_n(R_i) - F(R_i, \hat{\theta})\}^2, \tag{3}$$

where $F_n(\cdot)$ is the empirical distribution function based on $\{R_i\}_{i=1}^n$.

The asymptotic distributions of test statistics based on the empirical distribution when parameters are estimated have widely been investigated by [3], [8], [12], [4] and [11], between others. [3] considered the CVM test statistic when the null distribution function depends upon a nuisance parameter that must be estimated. [8] analysed the case when null distribution is the normal distribution with mean and variance both unknown. [12] extended the theory of the CVM test statistic to the case when the null distribution is an arbitrary $F(x, \theta)$, being θ a k -dimensional vector of unknown parameters.

A comprehensive treatment of the basic theory was given by [4]. [4] studied the weak convergence of a sample distribution function under a given sequence of alternative hypotheses when parameters are estimated from the data. His main result was that, in general, the asymptotic distribution of the test-statistic is not distribution-free on H_0 since its asymptotic distribution depends on F . Moreover, it is not even asymptotically parameter-free since this distribution depends in general on the value of θ . Consequently, the standard tables used for the traditional CVM test are not longer valid if some parameters are estimated from the sample. If these critical values are used in presence of nuisance parameters, the results will be conservative in the sense that the probability of a type I error will be smaller than as given by tables of the CVM test statistic. [11] provided percentage points for the asymptotic distribution of the CVM test statistic, for the cases where the distribution tested is (i) normal, with

mean or variance, or both, unknown; and (ii) exponential, with scale parameter unknown. However, percentage points for the asymptotic distribution of the CVM test statistic have not been obtained when the null distribution is different from the cases above.

Consequently, there are two alternatives to implement the test using W_n^2 . The first alternative is to derive asymptotic critical values from the limit distribution of W_n^2 and examine how $F(\cdot)$ and θ influence the results. The second alternative is to derive bootstrap critical values, designing an appropriate procedure and establishing its consistency.

The usefulness of bootstrap procedures in nonparametric distance tests was first highlighted by [10] and, since then, it has been extensively used in similar problems to ours. [1] showed that the bootstrap methodology consistently estimates the null distributions of various distance tests, included the CVM type test-statistics. They focused on parametric and nonparametric bootstrap procedures and they demonstrate that both procedures lead to correct asymptotic levels, that is, they lead to consistent estimates of the percentiles of the true limiting distribution of the test statistic when the parameters are estimated. However, in the case of nonparametric bootstrap, a correction of the bias is required.

4 Parametric Bootstrap Procedure and Monte Carlo Simulations

We test the null hypothesis that $R_i \equiv B(p, q)$, assuming that p and q are unknown. For random number generation we use GAUSS's 'rndbeta' command, which computes pseudo-random numbers with beta distribution. We choose the initial seed for the generator of 345567 and we introduce the GAUSS's command 'rndseed seed'. Otherwise, the default is that GAUSS uses the clock to generate an initial seed. The steps of the designed parametric bootstrap procedure are the following:

- *Step 1:* Assume R_1, R_2, \dots, R_n are i.i.d. random variables from a Beta distribution $B(p, q)$, with both p and q unknown. Under H_0 , estimate $\theta=(p, q)$ to obtain $\hat{\theta} = (\hat{p}, \hat{q})$. Evaluate the statistic W_n^2 using R_1, R_2, \dots, R_n and $\hat{\theta} = (\hat{p}, \hat{q})$.
- *Step 2:* Generate B bootstrap samples of i.i.d. observations $R^* = (R_1^*, R_2^*, \dots, R_n^*)'$ from $B(\hat{p}, \hat{q})$. Obtain $R_b^* = (R_{1b}^*, R_{2b}^*, \dots, R_{nb}^*)'$ for $b=1, \dots, B$. For each b , calculate new estimates $\hat{\theta}^* = (\hat{p}^*, \hat{q}^*)$. Compute the bootstrap version of W_n^2 , say W_{nb}^{2*} . In this way, a sample of B independent (conditionally on the original sample) observations of W_n^{2*} , say $W_{n1}^{2*}, \dots, W_{nB}^{2*}$, is obtained.
- *Step 3:* Let $W_{n(1-\alpha)B}^{2*}$ the $(1-\alpha)B$ -th order statistic of the sample $W_{n1}^{2*}, \dots, W_{nB}^{2*}$. Reject H_0 at the significance level α if $W_n^2 > W_{n(1-\alpha)B}^{2*}$. Additionally, the bootstrap p-value can be compute as $p_B = \text{card}(W_{nb}^{2*} \geq W_n^2) / B$.

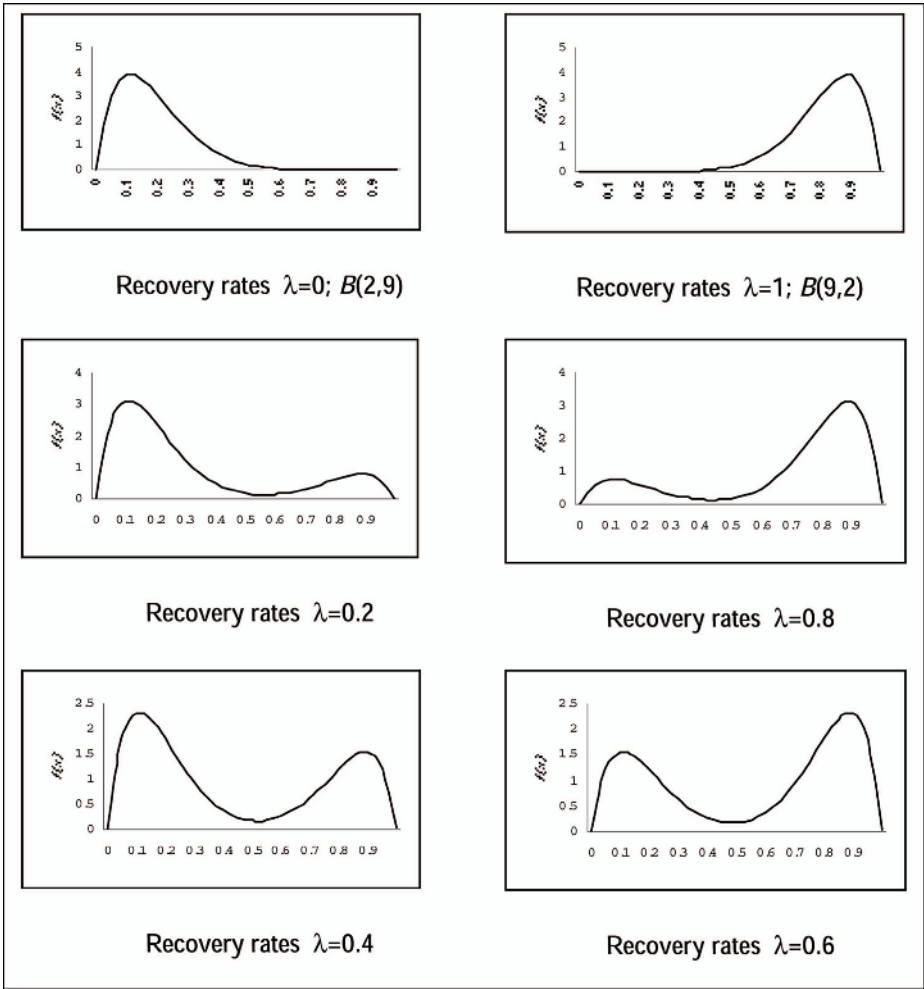


Fig. 1. This figure shows the shape of the pdf associated to a mixture of two Beta distributions, a $B(2,9)$ and a $B(9,2)$ with probability parameter λ . When $\lambda=0$ the resulting pdf is left-skewed shaped and, when $\lambda=1$, it is right-skewed shaped.

In order to check the accuracy of the bootstrap procedure, we perform some Monte Carlo experiments. In all cases we test H_0 at the 10%, 5% and 1% with the statistic W_n^2 . All experiments reported in this paper are conducted for different sample sizes $n=125, 250$ and 500 , with $B=500$ bootstrap replications. For each experiment we report the proportion of the rejections of H_0 based on 1000 simulation runs. To compute the test statistic W_n^2 , p and q are estimated by maximum likelihood procedure and by matching moments procedure.

In our procedure, we generate i.i.d. observations $R_i \equiv (1 - \lambda)B(2,9) + \lambda B(9,2)$, $i=1, \dots, n$, for various λ , $\lambda=0, 0.2, 0.4, 0.6, 0.8$ and 1 . Thus, H_0 is true if and only if $\lambda=0$ or 1 . In Figure 1, the corresponding pdf are plotted. As Figure 1 shows, a $B(2,9)$ and a $B(9,2)$ can represent the behavior of recovery rates under two different scenarios. On one hand, the pdf of a $B(2,9)$ is concentrated to the left and it could be feasible to catch the behavior of securities with mean recovery rate approximately equals to 18%, which represents an unsecured security. On the other hand, the probability density function of a $B(9,2)$ is concentrated to the right, which means that it could be reasonably modeled a secured security, with mean recovery rate approximately equals to 81%. Additionally, the mixture of two Beta distributions can describe the shape of the empirical distribution of recovery rates corresponding to portfolios composed by secured and unsecured securities. When $\lambda=0.2, 0.4, 0.6, 0.8$ the pdf is two-peaked. These cases are considered in order to generate data from different data generated processes that can not be assumed to follow a Beta distribution. For example, when $\lambda=0.2$, most of the probability mass is accumulated in the left-hand side of the distribution but there is a small peak in the left-hand side which makes this pdf slightly different from a Beta distribution.

Table 1. Proportion of Rejections of H_0 when the parameters p and q are estimated by maximum likelihood procedure

$X_i \sim (1-\lambda)B(2,9) + \lambda B(9,2), i=1, \dots, n$									
λ	$n=125$			$n=250$			$n=500$		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
0	10.4	4.7	1.4	8.2	4.4	1.2	9.6	4.1	0.9
0.2	64	52.3	30.1	93	86.3	67.4	99.6	99.1	96.1
0.4	46.8	35.6	15.6	67.4	55.6	31.7	91.9	85.9	68.4
0.6	49.1	37.3	18.9	68.9	58.2	33.9	92.1	87.3	70.2
0.8	63.7	50.2	28.9	90.4	84.1	65.1	99.8	99.2	96.3
1	9	4.9	1.4	9.2	5.1	1	10.8	5.9	1.3

Table 2. Proportion of Rejections of H_0 when the parameters p and q are estimated by matching moments procedure

$X_i \sim (1-\lambda)B(2,9) + \lambda B(9,2), i=1, \dots, n$									
λ	$n=125$			$n=250$			$n=500$		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
0	9.9	5.7	1.6	10.5	4.7	1.3	9.1	4.6	1.4
0.2	62.8	51.2	28.4	89.9	82.7	65.2	99.8	99.2	95.4
0.4	34.5	24.2	10.1	58.1	47.9	28	82.4	74.1	56.1
0.6	36.1	26.1	12.6	53.8	42.7	24.8	81.4	74	55.5
0.8	63.9	52.7	29.5	89.4	81.6	63.8	99.6	99.1	95.6
1	9.7	5.3	0.9	9.3	5.3	1.3	8.9	4.2	1

In Table 1 we report the results when the maximum likelihood estimation procedure is used while in Table 2 we show the results when the matching moments estimation procedure is chosen. In order to evaluate if the difference between the percentage of rejections of H_0 , r , and the nominal significance level α , we compute the Z statistic given by:

$$Z = \frac{(r - \alpha)}{\sqrt{\alpha(1 - \alpha) / R}} , \quad (4)$$

where R is the number of Monte Carlo replications and Z distributes normally with mean 0 and variance 1. In particular, the bootstrap test is well specified at the 95% level if $r \in [0.38, 1.61]$, $r \in [3.65, 6.35]$ and $r \in [8.14, 11.86]$ for $R=1000$ and, $\alpha=1\%$, 5% and 10%, respectively.

Our results show that the bootstrap test works reasonably well under both H_0 and H_1 , when the parameters p and q are estimated by maximum likelihood procedure or by matching moments procedure. The estimation method does not seem to affect the empirical size of the test. Under both estimation methods the empirical size is close to the nominal size of the test, as can be checked using the Z statistic. However, the bootstrap test under maximum likelihood estimation is slightly more powerful than when matching moments estimation is used.

In the results reported in Tables 1 and 2, there is evidence of power increase as the sample size arises. For the smallest sample size, $n=125$, the bootstrap test has quite low power, especially for the 1% level. This fact is not present at the larger sample sizes, suggesting that a sample size equals to 250 is enough to assure reasonably high power of the bootstrap test at 1% level. It is worth noting that the power of the bootstrap test decreases as λ gets nearer to 0.5. It is due to that the resulting mixture of beta distributions is U-shaped, which can be captured by a beta distribution.

5 Conclusions

In this paper, we have shown using computational methods how to test if the recovery rates follow a Beta distribution using a parametric bootstrap procedure to estimate the asymptotic null distribution of the Cramer-von Mises test statistic, when there are some unknown parameters that must be estimated from the sample. We have estimated the unknown parameters by the maximum likelihood procedure with general constraints on the parameters. For the maximization of the log-likelihood function we have chosen the Newton-Raphson numerical method in which both first and second derivative information are used. The implementation of the Newton-Raphson method involves a numerical calculation of the Hessian. Our simulations reveal that the bootstrap version of the Cramer-von Mises test statistic performs reasonably well under both H_0 and H_1 . Moreover, its performance is not hardly affected by the estimation method chosen to estimate the two parameters which characterize the Beta distribution. Since it is quite often necessary in credit risk analysis to test the true underlying distribution of recovery rates in order to estimate potential credit losses, our bootstrap procedure appears to be a reliable method to implement the test. Furthermore, our bootstrap procedure is flexible enough to be used to test any other possible null distri-

bution of interest. To sum up, computational methods should be taken into account in managing credit risk to ensure a reliable recovery rate distribution.

References

1. Babu, G.J., Rao, C.R.: Goodness-of-fit Tests when Parameters Are Estimated. *Sankhya: Indian J. Statist.* 66 (2004) 63-74
2. Carty, L.V., Lieberman, D.: *Corporate Bond Defaults and Default Rates*. Moody's Investor Service (1996)
3. Darling, D.A.: The Cramer-Smirnov Test in the Parametric Case. *Ann. Math. Statist.* 26 (1955) 1-20
4. Durbin, J.: Weak Convergence of the Sample Distribution Function when Parameters Are Estimated. *Ann. Statist.* 1 (1973) 279-290
5. Gordy, M., Jones, D.: *Capital Allocation for Securitizations with Uncertainty in Loss Prioritization*. Federal Reserve Board (2002)
6. Gupton, G.M., Stein, R.M.: *LossCalcTM: Model for Predicting Loss Given Default (LGD)*. Moody's Investors Service (2002)
7. Gupton, G.M., Stein, R.M.: *LossCalc2: Dynamic Prediction of LGD*. Moody's Investors Service (2005)
8. Kac, M., Kiefer, J., Wolfowitz, J.: On tests of Normality and other Tests of Goodness-of-fit, Based on Distance Method. *Ann. Math. Statist.* 26 (1955) 189-211
9. Pesaran, M.H., Schuermann, T., Treutler, B.J., Weiner, S.M.: *Macroeconomic Dynamics and Credit Risk: a Global Perspective*. DAE Working Paper No 0330 (2003) University of Cambridge
10. Romano, J.P.: A Bootstrap Revival of some Nonparametric Distance Tests. *J. Amer. Statist. Assoc.* 83 (1988) 698-708
11. Stephens, M.A.: Asymptotic Results for Goodness-of-fit Statistics with Unknown Parameters. *Ann. Statist.* 4 (1976) 357-369
12. Sukhatme, S.: Fredholm Determinant of a Positive Definite Kernel of a Special Type and its Application. *Ann. Math. Statist.* 43 (1972) 1914-1926

Using Performance Profiles to Evaluate Preconditioners for Iterative Methods

Michael Lazzareschi and Tzu-Yi Chen*

Department of Computer Science, Pomona College, Claremont CA 91711, USA
{md112002, tzuyi}@cs.pomona.edu

Abstract. We evaluate performance profiles as a method for comparing preconditioners for iterative solvers by using them to address three questions that have previously been asked about incomplete LU preconditioners. For example, we use performance profiles to quantify the observation that if a system can be solved by a preconditioned iterative solver, then that solver is likely to use less space, and not much more time, than a direct solver. In contrast, we also observe that performance profiles are difficult to use for choosing specific parameter values. We end by discussing the role performance profiles might eventually play in helping users choose a preconditioner.

Keywords: Iterative methods, preconditioners, performance profiles.

1 Introduction

Iterative methods for solving large, sparse linear systems $Ax = b$ are generally used when the time or memory requirements of a direct solver are unacceptable. However, getting good performance from an iterative method often requires first applying an appropriate preconditioner. Ideally, this preconditioner is inexpensive to compute and apply, and it generates a modified system $A'x' = b'$ that is somehow “better” for the iterative solver being used. Unfortunately, choosing an effective preconditioner for any given system is not usually a simple task.

This has led to the proposal of many different preconditioners, combined with an interest in helping users choose from among those preconditioners. For the latter, some researchers have suggested rules-of-thumb based on experimentation (eg, [7, 9]), whereas others have tried using data-mining techniques to extract general guidelines (eg, [24]). While the former can suggest default settings, it does not necessarily suggest when and how to adjust those defaults. While the latter could potentially provide more flexible guidelines, so far there has been limited success in extracting meaningful features. In this paper we consider using performance profiles [14], a technique introduced by the optimization community, to compare preconditioners for sparse linear systems.¹

* Corresponding author.

¹ For a survey of other methods used by the optimization community to compare heuristics, see [19].

Because performance profiles are known to be potentially misleading when used to compare very different solvers [15], in this initial study we only evaluate different value-based incomplete LU (ILU) preconditioners. We find that performance profiles can still be difficult to interpret even when restricted to this limited class of preconditioners. However, they do provide a way to begin addressing different questions about ILU preconditioners, including those asked in papers such as [7, 8, 18]. We end by discussing the role performance profiles might play in evaluating preconditioners.

2 Background

Value-based ILU preconditioners are a popular class of preconditioners which are often computed by mimicking the computation of a complete LU factorization, but also “dropping” small entries by setting their values to 0. Surveys such as [1, 5, 23] speak to the variety of preconditioners in this class; researchers have suggested different ways of deciding what elements to drop, different methods of ordering the matrix prior to computing the incomplete factorization, and other variations.

Given the variety of ILU preconditioners that exist, comparisons are necessary from a user’s point of view. Unfortunately, comparing different preconditioners can be difficult in part because of inherent trade-offs between the speed of convergence, memory requirements, and other attributes. Furthermore, different applications stress different criteria, with perhaps the only consensus being that preconditioned solvers that never converge to the correct solution, and preconditioned solvers that require both more space and more time than a direct method, are useless.

Currently, when new ILU preconditioners are proposed, or existing ones evaluated, authors often describe their behavior in tables where each row is a single matrix and each column is a measure such as the amount of space needed, the number of iterations, the accuracy to which the solution is computed, and so on (eg, [4, 6]). While these tables present complete data, they are inherently limited by the fact that users are unlikely to want to study a table spanning several pages. In addition, users remain responsible for determining what portions of the table are most relevant to their system. Other papers use graphs to visually display performance data (eg, [8, 18]). Although graphs can present data more compactly, they can also be difficult to interpret unless the author can point to clear trends. A generalizable method for presenting comparison data more compactly than tables, while still revealing essential differences along a number of dimensions, could help users in choosing a preconditioner.

We consider using performance profiles, introduced in [14], for this purpose. A performance profile comparing, say, the time needed to solve a set of systems S using any of a set of preconditioned solvers P , begins with the time taken for each preconditioned solver $p \in P$ to solve each system $s \in S$. From this timing data we compute $r_{p,s}$, which is the ratio of the time taken to solve s using the preconditioned solver p to the fastest time taken by any preconditioner to solve

s .² The performance profile for p over the sets P and S is a plot of the function $\rho_p(\tau)$, where

$$\rho_p(\tau) = \frac{|s \in S \text{ s.t. } r_{p,s} \leq \tau|}{|S|} \quad (1)$$

In other words, $\rho_p(\tau)$ is the probability that the solver p solves a problem in S in no more than τ times the minimum time taken by any solver in P .

Although performance profiles present less data than tables, the plots are potentially easier for users to interpret. In addition, performance profiles are sufficiently flexible that they could potentially be used as a standard method of comparison. However, this flexibility brings with it the potential for confusion: the appearance of a plot depends on the problems in S as well as on the solvers plotted in that graph. We ask to what extent these drawbacks impact the viability of using performance profiles as a general tool for evaluating preconditioners.

3 Methodology

We use a set of 115 nonsymmetric, real matrices taken primarily from the University of Florida sparse matrix collection [10]. The exact solution to the system $Ax = b$ was always assumed to be a vector of 1s, and the vector b was calculated accordingly. The iterative solver was GMRES(50) [22] with an upper bound of 500 on the number of iterations and a tolerance of 10^{-8} , choices which are also used in studies such as [7, 9].

We used the variant of ILU known as ILUTP_Mem [8], with values of `lfil` varying from 0 to 5, values of `pivtol` varying from 0 to 1, and values of `droptol` varying from 0 to .1. For each matrix and each combination of parameters we used the natural ordering as well as two fill-reducing orderings: RCM [20] and COLAMD [11]. We also experimented with MC64-based ordering and scaling for stability [16, 17]. To enable comparison to a direct solver, we ran SuperLU [13] on all the systems. Overall we ran several thousand test cases.

Because our focus in this paper is on understanding the kind of information that can be read from performance profiles, we discuss only a subset of our data. In particular, we always use `pivtol`= 1.0 and `droptol`= 0.0. We always use MC64(5) with scaling for stability, and any subsequent fill-reducing ordering is always applied symmetrically. Furthermore, although we collected a wide range of information about each run, in this paper we use only the overall time taken by the preconditioned solver (including the time to compute and apply the preconditioner, to run the iterative solver, and to recover the solution to the original problem) and the space used by the preconditioner.

We plot performance profiles comparing space and time, as these are two aspects of interest to most users. The profiles are plotted for values of τ up to 3 for space and up to 10 for time. The difference reflects the observation that

² Note that $r_{p,s} = 1.0$ if and only if p is the best solver for s , and $r_{p,s} = \infty$ if and only if s cannot be solved using p (ie, the iterative solver did not converge).

users can typically wait longer for a solution if necessary, but that excessive space requirements are insurmountable.

Finally, in the legend for each performance profile we note the total percentage of problems that can be solved with each solver. In this way the focus of the graphs remains on small values of τ , which represent the regime where the solvers are most competitive. However, by giving the overall percentage of problems solved, we also reveal whether a relatively inefficient solver might be particularly robust.

4 Examples and Analysis

In this section we use performance profiles to address three questions of interest when evaluating ILU preconditioners. As noted in [15], and as will become evident here, performance profiles can be confusing when too many, or very different, solvers are plotted in a single graph. Hence we restrict each profile in this section to a relatively small set of preconditioned iterative solvers.

4.1 Comparing Fill-Reducing Orderings

Just as fill-reducing orderings are used to reduce the fill in complete LU factorizations, they can also be used with ILU factorizations to compute more accurate preconditioners in less space. A general rule of thumb that emerges from previous work is that minimum degree methods are better than other fill-reducing orderings for value-based ILU preconditioners (eg, [2, 3, 7, 12, 18]). We use performance profiles to reevaluate this claim.

Figure 1 compares natural, RCM, and COLAMD orderings on their time and requirements in the graphs on the left and right, respectively. We use an

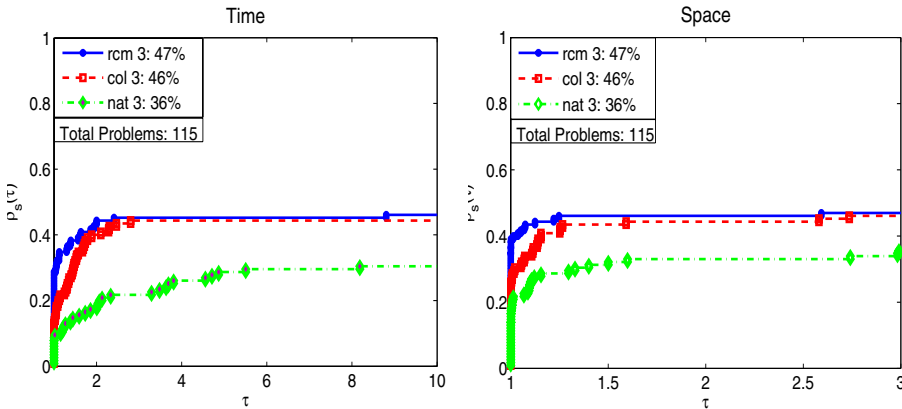


Fig. 1. Performance profiles comparing time and space requirements for ILUTP_Mem preconditioned GMRES(50) using three different orderings: natural, RCM, and COLAMD.

`lfil` value of 3 for these examples based on results in [8] that suggest higher levels of fill do not significantly increase the likelihood of convergence for the `ILUTP_Mem` preconditioner.

The two graphs show that performance is similar using either the RCM or COLAMD orderings, and that both outperform the natural ordering. The percentages in the legends tell us that RCM solves more problems overall. Furthermore, the fact that the lines for RCM are always above those for COLAMD shows that the former solves more problems within any given percentage of the overall best time or space. However, the steep slope between $\tau = 1$ and $\tau = 2$ for COLAMD in the time plot indicates that COLAMD rarely takes more than twice as much time as the best of the three orderings.

Clearly COLAMD is not significantly better than RCM as a fill-reducing ordering for use with this particular ILU preconditioner and these parameter settings. Further experiments might reveal whether COLAMD is, as suggested by previous work, notably better in other contexts.

4.2 Choosing the Value of `lfil`

Value-based ILU preconditioners often provide parameters for controlling the trade-off between space and time. Typically, the more space used by a preconditioner, the fewer iterations (and hopefully less overall time) needed for the subsequent iterative solver to converge. In ILUT [21] and its variants, the `lfil` parameter gives one way to control this trade-off.³ To better understand the effect of the value of `lfil`, Figure 2 plots performance profiles for time and space as a function of `lfil` values ranging from 0 to 5. Based on the results in the previous section, RCM ordering is used.

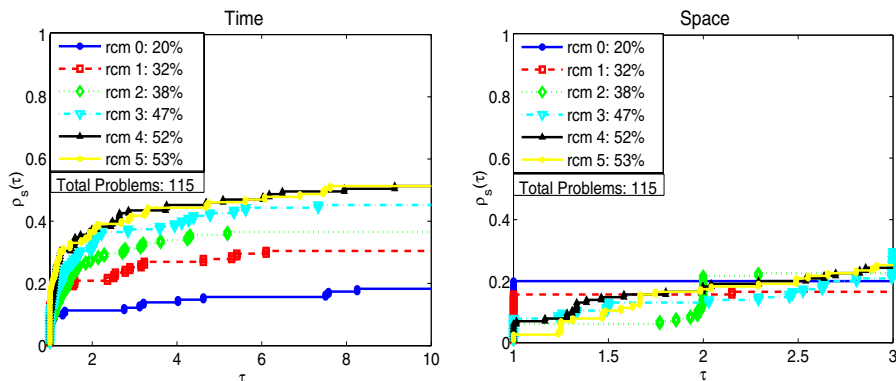


Fig. 2. Performance profiles comparing time and space requirements as a function of different values of `lfil`

³ The original meaning of the `lfil` parameter in ILUT is found to be potentially misleading in [8]; we use the modified definition of the `lfil` parameter they suggest.

The fact that these profiles are more complex than those in the previous section reinforces the observation that plotting more than a few solvers on one plot is inadvisable. The plot on the left seems to indicate that a higher `lfil` value leads to a greater likelihood of convergence within any small multiple of the optimal time. However, since many more problems can be solved using larger values of `lfil`, as indicated by the percentages in the legend, this observation is not very insightful. At most we can say that once τ is larger than about 1.5, higher `lfil` values seem to be generally better, though the difference between `lfil`= 4 and 5 seems small. In other words, higher `lfil` values are generally better than lower values if our primary concern is with having a solver that is no more than 50% slower than optimal over `lfil` values between 0 and 5.

The performance profile for space in Figure 2 is also difficult to interpret. As expected, lower `lfil` values generally use less space, but the magnitude and consistency of the advantage are unclear. This is due to the fact that, in general, if a system can be solved with `lfil`= k , it can also be solved with `lfil`> k , albeit using more space. This, and the fact that the ILUTP_Mem preconditioner uses about twice as much space with `lfil`= 2 as it does with `lfil`= 1, explains the jumps at integer values of τ . In particular, the fact that a problem which can be solved with `lfil`= 1 can also be solved with `lfil`= 2 is not reflected until $\tau = 2$, when we plot all solvers that can solve with up to twice as much space as optimal.

While the above explain some of the trends exhibited in the plots in Figure 2, it does not provide clear insight into how to set `lfil` to optimize any meaningful measure of performance.

4.3 Evaluating Robustness

Finally we address a question that is sometimes overlooked when evaluating preconditioners for iterative solvers: when, if ever, are they competitive with direct solvers? Figure 3 shows performance profiles for time and space comparing

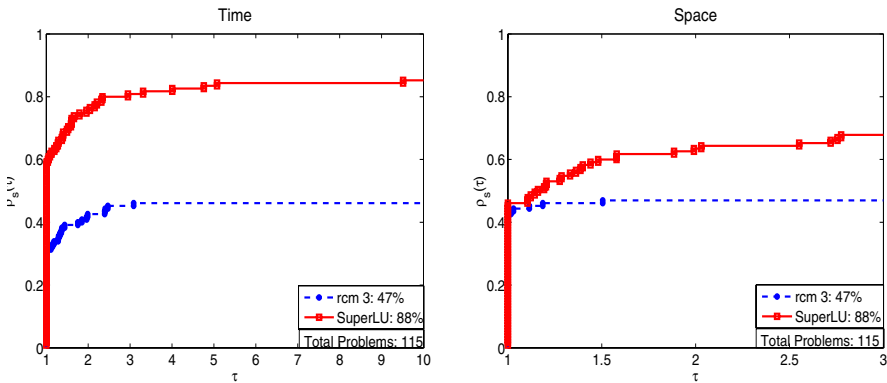


Fig. 3. Performance profiles comparing time and space for RCM-ordered ILUTP_Mem and SuperLU on the full set of 115 matrices

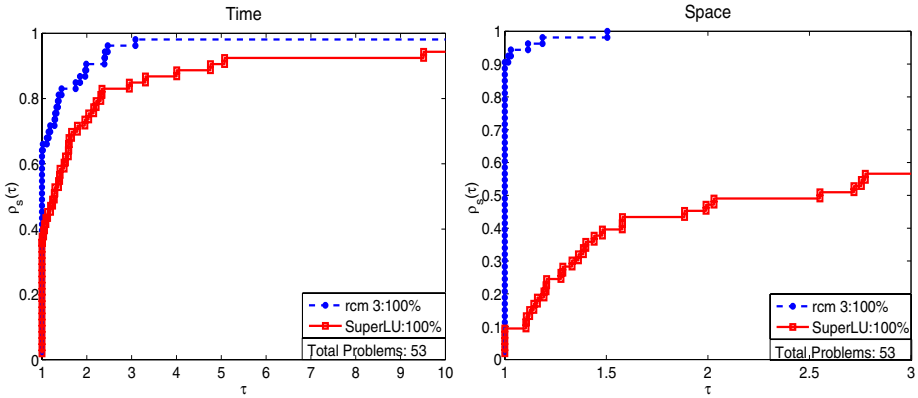


Fig. 4. Performance profiles comparing time and space for RCM-ordered ILUTP_Mem and SuperLU on the subset of 53 systems which they can both solve

the direct solver SuperLU [13] and RCM-ordered ILUTP_Mem preconditioned GMRES(50), with `lfil=3`.

Initially SuperLU looks far superior. However, as previously noted, the fact that it converges on many more systems complicates the interpretation of the performance profiles. If we note that SuperLU is clearly more robust and then replot the performance profiles only on the subset of the problems that are solved by both methods, a different picture emerges.

In particular, Figure 4 shows that the preconditioned iterative solver uses less space and that it never uses more than about 50% more space than that used by the more space-efficient of the two solvers. Furthermore, the preconditioned iterative solver is rarely more than three times slower than the faster of the two solvers. From this we conclude that if one has reason to believe a particular system can be solved using an iterative method, then the potential benefits in both time and space are well worth considering.

5 Conclusion

In this paper we discuss the potential use of performance profiles in evaluating ILU preconditioned iterative solvers. Regarding the use of performance profiles for this purpose, we find that they can be a useful way of presenting a summary of results over a large number of systems. However, they can be difficult to interpret when used to compare too many, or very different, solvers in a single plot. We also observe that plotting these profiles only for small values of τ , and additionally noting the total percentage of problems on which each solver converges to give a sense of behavior at large values of τ , makes these plots easier to interpret. From a user perspective, we find that if a user believes their a system can be solved by a preconditioned iterative method, then that solver will likely be both space and time competitive with direct methods. This suggests further work into understanding what makes a system solvable by preconditioned iterative methods.

Taken as a whole, the above suggest that what might be useful would be a tool that allowed users to specify the exact solvers and the exact systems that they wanted to compare. While large tables of data such as those in some existing papers on preconditioners technically allow a user to do this, a standardized graph-based format might simplify the task of looking through data for insight into the behavior of different solvers on different classes of systems. While performance profiles might be useful in this role, a user would need to be given hints on how to meaningfully interpret these plots. We are still evaluating the strengths and weaknesses of performance profiles, and we are also interested in evaluating other methods for presenting comparison data.

Acknowledgements

This work was funded in part by the National Science Foundation under grant #CCF-0446604. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. M. Benzi. Preconditioning techniques for large linear systems: A survey. *J. of Comp. Physics*, 182(2):418–477, November 2002.
2. M. Benzi, W. D. Joubert, and G. Mateescu. Numerical experiments with parallel orderings for ILU preconditioners. *Electronic Transactions on Numerical Analysis*, pages 88–114, 1999.
3. M. Benzi, D. B. Szyld, and A. van Duin. Orderings for incomplete factorization preconditioning of nonsymmetric problems. *SIAM J. Sci. Comput.*, 20(5):1652–1670, 1999.
4. M. Benzi and M. Tuma. A sparse approximate inverse preconditioner for non-symmetric linear systems. *SIAM Journal on Scientific Computing*, 19:968–994, 1998.
5. T. F. Chan and H. A. van der Vorst. Approximate and incomplete factorizations. In D. E. Keyes, A. Samed, and V. Venkatakrishnan, editors, *Parallel numerical algorithms*, volume 4 of *ICASE/LaRC Interdisciplinary Series in Science and Engineering*, pages 167–202. Kluwer Academic, Dordrecht, 1997.
6. A. Chapman, Y. Saad, and L. Wigton. High-order ILU preconditioners for CFD problems. *Int. J. Numer. Meth. Fluids*, 33:767–788, 2000.
7. T.-Y. Chen. *Preconditioning sparse matrices for computing eigenvalues and solving linear systems of equations*. PhD thesis, University of California at Berkeley, December 2001.
8. T.-Y. Chen. ILUTP_Mem: A space-efficient incomplete LU preconditioner. In A. Laganà, M. L. Gavrilova, V. Kumar, Y. Mun, C. J. K. Tan, and O. Gervasi, editors, *Proceedings of the 2004 International Conference on Computational Science and its Applications*, volume 3046 of *LNCS*, pages 31–39, 2004.
9. E. Chow and Y. Saad. Experimental study of ILU preconditioners for indefinite matrices. *J. Comp. and Appl. Math.*, 86:387–414, 1997.

10. T. Davis. University of Florida sparse matrix collection. NA Digest, v.92, n.42, Oct. 16, 1994 and NA Digest, v.96, n.28, Jul. 23, 1996, and NA Digest, v.97, n.23, June 7, 1997. Available at: <http://www.cise.ufl.edu/research/sparse/matrices/>.
11. T. Davis, J. Gilbert, S. Larimore, and E. Ng. A column approximate minimum degree ordering algorithm. *ACM Trans. on Math. Softw.*, 30(3):353–376, September 2004.
12. E. F. D’Azevedo, P. A. Forsyth, and W.-P. Tang. Ordering methods for preconditioned conjugate gradient methods applied to unstructured grid problems. *SIAM J. Matrix Anal. Appl.*, 13(3):944–961, July 1992.
13. J. W. Demmel, J. R. Gilbert, and X. S. Li. *SuperLU users’ guide*, September 1999. Available at: <http://www.nersc.gov/~xiaoye/SuperLU/>.
14. E. Dolan and J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91:201–213, 2002.
15. E. Dolan, J. Moré, and T. Munson. Optimality measures for performance profiles. *Preprint ANL/MCS-P1155-0504*, 2004.
16. I. S. Duff and J. Koster. The design and use of algorithms for permuting large entries to the diagonal of sparse matrices. *SIAM J. Matrix Anal. Appl.*, 20(4):889–901, 1999.
17. I. S. Duff and J. Koster. On algorithms for permuting large entries to the diagonal of a sparse matrix. *SIAM J. Matrix Anal. Appl.*, 22(4):973–996, 2001.
18. J. R. Gilbert and S. Toledo. An assessment of incomplete-LU preconditioners for nonsymmetric linear systems. *Informatica*, 24:409–425, 2000.
19. C. Khompatraporn, J. D. Pinter, and Z. B. Zabinsky. Comparative assessment of algorithms and software for global optimization. *J. Global Opt.*, 31(4):613–633, 2005.
20. W.-H. Liu and A. H. Sherman. Comparative analysis of the Cuthill-McKee and the reverse Cuthill-McKee ordering algorithms for sparse matrices. *SIAM J. Num. Anal.*, 13(2):198–213, April 1976.
21. Y. Saad. ILUT: A dual threshold incomplete LU factorization. *Numer. Linear Algebra Appl.*, 4:387–402, 1994.
22. Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 7(3):856–869, July 1986.
23. Y. Saad and H. A. van der Vorst. Iterative solution of linear systems in the 20th century. *J. of Comp. and Appl. Math.*, 123:1–33, November 2000.
24. S. Xu and J. Zhang. Solvability prediction of sparse matrices with matrix structure-based preconditioners. In *Proceedings of Preconditioning 2005*, Atlanta, Georgia, May 2005.

Multicast ω -Trees Based on Statistical Analysis*

Moonseong Kim¹, Young-Cheol Bang², and Hyunseung Choo¹

¹ School of Information and Communication Engineering,
Sungkyunkwan University, 440-746, Suwon, Korea
Tel.: +82-31-290-7145

{moonseong, choo}@ece.skku.ac.kr
² Department of Computer Engineering,
Korea Polytechnic University, 429-793, Gyeonggi-Do, Korea
Tel.: +82-31-496-8292
ybang@kpu.ac.kr

Abstract. In this paper, we study the efficient multicast routing tree problem with QoS requirements. The new multicast weight parameter is proposed by efficiently combining two independent measures, the link cost and delay. The weight $\omega \in [0, 1]$ plays an important role in combining the two measures. If the ω approaches to 0, then the tree delay is decreasing. On the contrary if it closes to 1, the tree cost is decreasing. Therefore, if the ω is decided, then the efficient multicast tree can be found. A case study shows various multicast trees for each ω . When network users have various QoS requirements, the proposed multicast weight parameter is very informative for them.

1 Introduction

Multicast Services have been increasingly used by various continuous media applications. For example, the multicast backbone (MBONE) of the Internet has been used to transport real time audio and video for news, entertainment, and distance learning. In multicast communications, messages are sent to multiple destinations that belong to the same multicast group. These group applications demand a certain amount of reserved resources to satisfy their Quality of Service (QoS) requirements such as end-to-end delay, delay jitter, loss, cost, throughputs, and etc. Since resources for multicast tree are reserved along a given path to each destination in a given multicast tree, it may fail to construct a multicast tree to guarantee the required QoS if a single link cannot support required resources. Thus an efficient solution for multicast communications includes the construction of a multicast tree that has the best chance to satisfy the resource requirements [1, 2, 8, 9, 10, 11, 17, 19, 21].

* This research was supported by the Ministry of Information and Communication, Korea under the Information Technology Research Center support program supervised by the Institute of Information Technology Assessment, IITA-2005-(C1090-0501-0019) and the Ministry of Commerce, Industry and Energy under Next-Generation Growth Engine Industry. Dr. Choo is the corresponding author and Dr. Bang is the co-corresponding author.

Previous optimization techniques for multicast routing have considered two optimization goals, delay optimization and cost optimization, but as distinct problems. The optimal delay solution is such that the sum of delays on the links along the path from source to each destination is minimum. Dijkstra's shortest path algorithm [4] can be used to generate the shortest paths from the source to the destination nodes in $O(n^2)$ time in a graph with n nodes. This provides the optimal solution for delay optimization. A cost optimized multicast route is a tree spanning the destinations such that the sum of the costs on the links of the tree is minimum. This problem is also known as the Steiner tree problem [9], and is known to be NP-complete [7]. However, some heuristics for the Steiner tree problem have been developed that take polynomial time [11,17] and produce near optimal results.

In this paper, we consider multicast routing as a source routing problem, with each node having full knowledge of the network and its status. We associate a link cost and a link delay with each link in the network. The problem is to construct a tree spanning the destination nodes, such that it has the QoS requirements such as minimum cost tree, minimum delay tree, cost-constrained minimum delay tree, delay-bounded minimum cost tree problem, and etc. The cost of optimal delay solution is relatively more expensive than the cost of cost optimized multicast route, and moreover, the delay of the cost optimized multicast route is relatively higher than the delay of optimal delay solution. The negotiation between the tree cost and the tree delay is important. Hence, we introduce the new multicast parameter that regulates both the cost and the delay at the same time. We use TM algorithm [17], well-known Steiner tree algorithm, with the proposed parameter and can adjust them by the weight $\omega \in [0, 1]$.

The rest of paper is organized as follows. In Section 2, we describe the network model and a well known TM algorithm. Section 3 presents details of the new parameter and illustrates with example. Then we analyze and evaluate the performance of the proposed parameter by simulation in Section 4. Section 5 concludes this paper.

2 Preliminaries

2.1 Network Model

We consider that a computer network is represented by a directed graph $G = (V, E)$ with n nodes and l links, where V is a set of nodes and E is a set of links, respectively. Each link $e = (i, j) \in E$ is associated with two parameters, namely link cost $c(e) \geq 0$ and link delay $d(e) \geq 0$. The delay of a link, $d(e)$, is the sum of the perceived queueing delay, transmission delay, and propagation delay. We define a path as sequence of links such that $(u, i), (i, j), \dots, (k, v)$, belongs to E .

Let $P(u, v) = \{(u, i), (i, j), \dots, (k, v)\}$ denote the path from node u to node v . If all nodes u, i, j, \dots, k, v are distinct, then we say that it is a simple directed path. We define the length of the path $P(u, v)$, denoted by $n(P(u, v))$, as a number of links in $P(u, v)$. For given a source node $s \in V$ and a destination node $d \in V$, $(2^{s \rightarrow d}, \infty)$ is the set of all possible paths from s to d .

$$(2^{s \rightarrow d}, \infty) = \{ P_k(s, d) \mid \text{all possible paths from } s \text{ to } d, \forall s, d \in V, \forall k \in \Lambda \}$$

where Λ is an index set. The path cost of P_k is given by $\phi_C(P_k) = \sum_{e \in P_k} c(e)$ and the path delay of P_k is given by $\phi_D(P_k) = \sum_{e \in P_k} d(e), \forall P_k \in (2^{s \rightarrow d}, \infty)$.

For the multicast communications, messages need to be delivered to all receivers in the set $M \subseteq V \setminus \{s\}$ which is called the multicast group, where $|M| = m$. The path traversed by messages from the source s to a multicast receiver, m_i , is given by $P(s, m_i)$. Thus multicast routing tree can be defined as $T(s, M) = \bigcup_{m_i \in M} P(s, m_i)$ and the messages are sent from s to M through $T(s, M)$. The tree cost of tree $T(s, M)$ is given by $\phi_C(T(s, M)) = \sum_{e \in T} c(e)$ and the tree delay is $\phi_D(T(s, M)) = \max\{\phi_D(P(s, m_i)) \mid \forall P(s, m_i) \subseteq T, \forall m_i \in M\}$.

2.2 TM Algorithm

There is a well known approach for constructing multicast tree with minimum cost. The algorithm TM due to Takahashi and Matsuyama [17] is a shortest path based algorithm and works on asymmetric directed networks. Also it was further studied and generalized by Ramanathan [15]. The TM algorithm is very similar to the shortest path based Prim's minimum spanning tree algorithm [14], and works as following three steps.

1. Construct a subtree, T_1 , of network G , where T_1 consists a source node s only. Let $i = 1$ and $M_i = \{s\}$
2. Find the closest node, $m_i \in (M \setminus M_i)$, to T_i (if tie, broken arbitrarily). Construct a new subtree, T_{i+1} by adding all links on the minimum cost path from T_i to m_i . Set $M_i = M_i \cup \{m_i\}$. Set $i = i + 1$.
3. If $|M_i| < |M|$ then go to step 2, otherwise return a final T_i

Fig. 1 (a) shows a given network topology with link costs specified on each link. Fig. 1 (b) represents the ultimate multicast tree obtained by the TM. The cost of the tree generated by the TM is 13. In the worst case, the cost of tree by TM is worse than $2(1 - 1/|M|)\phi_C(T)$, where T is the optimal tree [17].

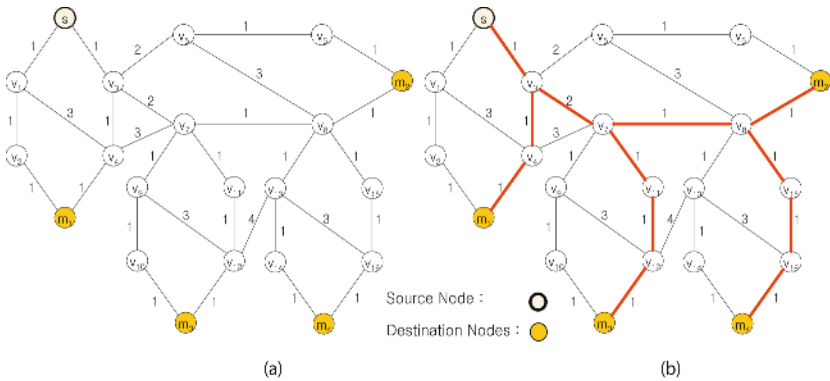


Fig. 1. Given a network (a), a multicast tree based on TM is shown in (b)

3 Proposed Multicast ω -Trees

3.1 The Negotiation Between Cost and Delay

We compute two paths the Least Delay path, $P_{LD}(s, m_k)$, and the Least Cost path, $P_{LC}(s, m_k)$, from the source s to the each destination m_k . Since only link-delays are considered to find $P_{LD}(s, m_k)$, $\phi_C(P_{LD})$ is always greater than or equal to $\phi_C(P_{LC})$. If the path cost, $\phi_C(P_{LD})$, is decreased by $100(1 - \frac{\phi_C(P_{LC})}{\phi_C(P_{LD})})\%$, the decreased value is obviously equal to $\phi_C(P_{LC})$.

The sample set, P_{LC} or P_{LD} from $(2^{s-d}, \infty)$, with non-normal distribution is approximately normally distributed by the Central Limit Theorem (CLT) [13]. The confidence interval $100(1 - \alpha)\%$ for the sample mean of the arbitrary sample set can be described by ‘‘Confidence Interval’’ in Fig. 2. Let \bar{C} be the average of link costs along P_{LD} with $(i, j) \in P_{LD}$ then $\bar{C} = \phi_C(P_{LD})/n(P_{LD})$. We employ the Gaussian distribution by the Central Limit Theorem. We consider the confidence interval $2 \times 100(1 - \frac{\phi_C(P_{LC})}{\phi_C(P_{LD})})\%$ to decrease $100(1 - \frac{\phi_C(P_{LC})}{\phi_C(P_{LD})})\%$ and should calculate its percentile. Since the Gaussian density function is symmetric to the mean \bar{C} , if the value that has to be decreased is greater than or equal to 50% then we interpret this value as 99.9% confidence interval for simplicity.

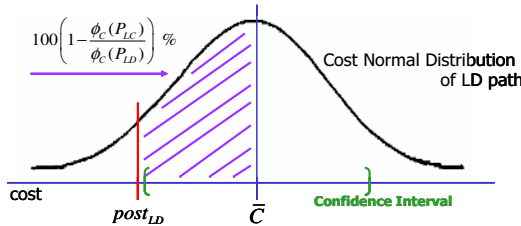


Fig. 2. $post_{LD}$

As shown in Fig. 2, $post_{LD}$ is the detection point to change the path cost from s to m_k . So, it is essential to find the percentile $z_{\alpha/2}$. In order to obtain it, we can use the cumulative distribution function (CDF). Let the CDF be $F(x)$ such that $\int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$. Then the percentile, $z_{\alpha/2}^d$, is a solution of the following equation, $F(z_{\alpha/2}^d) - \frac{1}{2} = 1 - \frac{\phi_C(P_{LC})}{\phi_C(P_{LD})}$ if $100(1 - \frac{\phi_C(P_{LC})}{\phi_C(P_{LD})})\% < 50\%$.

After calculating the percentile, we compute the $post_{LD} = \bar{C} - z_{\alpha/2}^d \frac{S_{LD}}{\sqrt{n(P_{LD})}}$ where S_{LD} , $(\sum_{e \in P_{LD}} (c(e) - \bar{C})^2 / (n(P_{LD}) - 1))^{1/2}$, is the sample standard deviation. If $n(P_{LD}) = 1$, then $S_{LD} = 0$. The new cost value of each link for m_k is as follow,

$$Cfct(e) = \max\{ 1, 1 + (c(e) - post_{LD}) \frac{\omega}{0.5} \}, 0 \leq \omega \leq 1 .$$

Meanwhile, $P_{LC}(s, m_k)$ is computed by taking the link-cost only into account. So, $\phi_D(P_{LC})$ is always greater than or equal to $\phi_D(P_{LD})$. If $\phi_D(P_{LC})$ is decreased by $100(1 - \frac{\phi_D(P_{LD})}{\phi_D(P_{LC})})\%$, then the decreased value is to be $\phi_D(P_{LD})$. The new delay value of each link for m_k can be derived by the same manner used in the case of P_{LD} ,

$$Dfct(e) = \max\{ 1, 1 + (d(e) - post_{LC}) \frac{1 - \omega}{0.5} \}, 0 \leq \omega \leq 1 .$$

Once the $Cfct(e)$ and the $Dfct(e)$ are computed, we calculate the new parameter values with weight ω , $Cfct(e) \times Dfct(e)$, for each link $e \in E$ of G for $m_k \in M$. Let $X^{m_k} = (Cfct \cdot Dfct)_{n \times n}$ be a adjacency matrix for $\forall m_k \in M$. And we normalize $X^{m_k} = (x_{ij}^{m_k})_{n \times n}$, it is called N^{m_k} .

$$i.e., N^{m_k} = (x_{ij}^{m_k})_{n \times n} / \max\{x_{ij}^{m_k} \mid 1 \leq i, j \leq n\}, \forall m_k \in M.$$

Finally, we obtain the new parameter, $N = \sum_{m_k \in M} N^{m_k}$, that regulates both the tree cost and the tree delay at the same time by the weight $\omega \in [0, 1]$. We use the TM algorithm [17] with the new weight parameter N . The weight ω plays an important role in combining the two measures, the delay and cost. If the ω approximates to 0, then the tree delay is reduced. Otherwise, it approaches to 1, and the tree cost is reduced. Therefore if a ω is decided, then the efficient routing tree can be found.

3.2 A Case Study

Fig. 3 shows a example of the trees obtained by different routing algorithms to span destinations. Fig. 3 (a) shows a given network topology G . Link costs and link delays are shown to each link as a pair $(cost, delay)$. To construct a tree rooted at v_0 that spans the destination set $M = \{v_2, v_3, v_4\}$, we consider either link cost or link delay. Fig. 3 (b) and (f) explain the optimal delay solution using Dijkstra’s algorithm and the cost optimized multicast route using TM algorithm, respectively. Fig. 3 (c)-(e) are good examples to illustrate multicast ω -Trees. In particular, we consider $\omega = 0.5$ for description. The following steps explain processes for obtaining new parameter N .

Firstly, we think the destination node v_2 . $P_{LD}(v_0, v_2) = \{(v_0, v_1), (v_1, v_4), (v_4, v_3), (v_3, v_2)\}$ and $P_{LC}(v_0, v_2) = \{(v_0, v_1), (v_1, v_2)\}$. $\phi_C(P_{LD}) = 19$, $\phi_C(P_{LC}) = 10$, $\phi_D(P_{LC}) = 16$, and $\phi_D(P_{LD}) = 13$. $\bar{C} = 19/4 = 4.75$ and $\bar{D} = 16/2 = 8$. $S_{LD} = \sqrt{14.92} = 3.86$ and $S_{LC} = \sqrt{8} = 2.83$. $100(1 - \frac{10}{19}) = 47.37\%$, so $z_{\alpha/2}^d = 1.88$ and $100(1 - \frac{13}{16}) = 18.75\%$, so $z_{\alpha/2}^c = 0.50$. $post_{LD} = 4.75 - 1.88 \frac{3.86}{\sqrt{4}} = 1.12$ and $post_{LC} = 8 - 0.50 \frac{2.83}{\sqrt{2}} = 7.00$. For a link (v_0, v_1) in Fig. 3 (d), we calculate $Cfct((v_0, v_1)) = \max\{1, 1 + (5 - 1.12) \frac{0.5}{0.5}\} = 4.88$ and $Dfct((v_0, v_1)) = \max\{1, 1 + (6 - 7) \frac{(1 - 0.5)}{0.5}\} = 1$. $Cfct((v_0, v_1)) \times Dfct((v_0, v_1)) = 4.88$. By the same manner, we obtain all new values in the network G for the destination node v_2 .

$$\begin{aligned}
 C^{v_2} &= \left(C_{fct}(e) \right)_{6 \times 6} = \begin{pmatrix} \cdot & 4.88 & \cdot & \cdot & \cdot & 6.88 \\ 4.88 & \cdot & 4.88 & 5.88 & 1.00 & 6.88 \\ \cdot & 4.88 & \cdot & 2.88 & \cdot & \cdot \\ \cdot & 5.88 & 2.88 & \cdot & 9.88 & \cdot \\ \cdot & 1.00 & \cdot & 9.88 & \cdot & 9.88 \\ 6.88 & 6.88 & \cdot & \cdot & 9.88 & \cdot \end{pmatrix}_{6 \times 6} \\
 D^{v_2} &= \left(D_{fct}(e) \right)_{6 \times 6} = \begin{pmatrix} \cdot & 1.00 & \cdot & \cdot & \cdot & 4.00 \\ 1.00 & \cdot & 4.00 & 1.00 & 1.00 & 1.00 \\ \cdot & 4.00 & \cdot & 1.00 & \cdot & \cdot \\ \cdot & 1.00 & 1.00 & \cdot & 1.00 & \cdot \\ \cdot & 1.00 & \cdot & 1.00 & \cdot & 1.00 \\ 4.00 & 1.00 & \cdot & \cdot & 1.00 & \cdot \end{pmatrix}_{6 \times 6} \\
 X^{v_2} &= \left(C_{fct}(e) \cdot D_{fct}(e) \right)_{6 \times 6} = \begin{pmatrix} \cdot & 4.88 & \cdot & \cdot & \cdot & 27.52 \\ 4.88 & \cdot & 19.52 & 5.88 & 1.00 & 6.88 \\ \cdot & 19.52 & \cdot & 2.88 & \cdot & \cdot \\ \cdot & 5.88 & 2.88 & \cdot & 9.88 & \cdot \\ \cdot & 1.00 & \cdot & 9.88 & \cdot & 9.88 \\ 27.52 & 6.88 & \cdot & \cdot & 9.88 & \cdot \end{pmatrix}_{6 \times 6} \\
 N^{v_2} &= \max\{x_{ij}^{v_2}\}^{-1} \left(x_{ij}^{v_2} \right)_{6 \times 6} = \begin{pmatrix} \cdot & 0.18 & \cdot & \cdot & \cdot & 1.00 \\ 0.18 & \cdot & 0.71 & 0.21 & 0.04 & 0.25 \\ \cdot & 0.71 & \cdot & 0.10 & \cdot & \cdot \\ \cdot & 0.21 & 0.10 & \cdot & 0.36 & \cdot \\ \cdot & 0.04 & \cdot & 0.36 & \cdot & 0.36 \\ 1.00 & 0.25 & \cdot & \cdot & 0.36 & \cdot \end{pmatrix}_{6 \times 6}
 \end{aligned}$$

Moreover, we think others destinations v_3 and v_4 .

$$\begin{aligned}
 N^{v_3} &= \begin{pmatrix} \cdot & 0.17 & \cdot & \cdot & \cdot & 1.00 \\ 0.17 & \cdot & 0.60 & 0.14 & 0.04 & 0.47 \\ \cdot & 0.60 & \cdot & 0.04 & \cdot & \cdot \\ \cdot & 0.14 & 0.04 & \cdot & 0.29 & \cdot \\ \cdot & 0.04 & \cdot & 0.29 & \cdot & 0.29 \\ 1.00 & 0.47 & \cdot & \cdot & 0.29 & \cdot \end{pmatrix}_{6 \times 6} \\
 N^{v_4} &= \begin{pmatrix} \cdot & 0.26 & \cdot & \cdot & \cdot & 1.00 \\ 0.26 & \cdot & 0.60 & 0.23 & 0.03 & 0.57 \\ \cdot & 0.60 & \cdot & 0.03 & \cdot & \cdot \\ \cdot & 0.23 & 0.03 & \cdot & 0.23 & \cdot \\ \cdot & 0.03 & \cdot & 0.23 & \cdot & 0.46 \\ 1.00 & 0.57 & \cdot & \cdot & 0.46 & \cdot \end{pmatrix}_{6 \times 6}
 \end{aligned}$$

Therefore, we obtain the new parameter N for a multicast tree $T_{\omega:0.5}$.

$$N = \sum_{m_k \in M} N^{m_k} = \begin{pmatrix} \cdot & 0.61 & \cdot & \cdot & \cdot & 3.00 \\ 0.61 & \cdot & 1.91 & 0.58 & 0.10 & 1.29 \\ \cdot & 1.91 & \cdot & 0.17 & \cdot & \cdot \\ \cdot & 0.58 & 0.17 & \cdot & 0.87 & \cdot \\ \cdot & 0.10 & \cdot & 0.87 & \cdot & 1.10 \\ 3.00 & 1.29 & \cdot & \cdot & 1.10 & \cdot \end{pmatrix}_{6 \times 6}$$

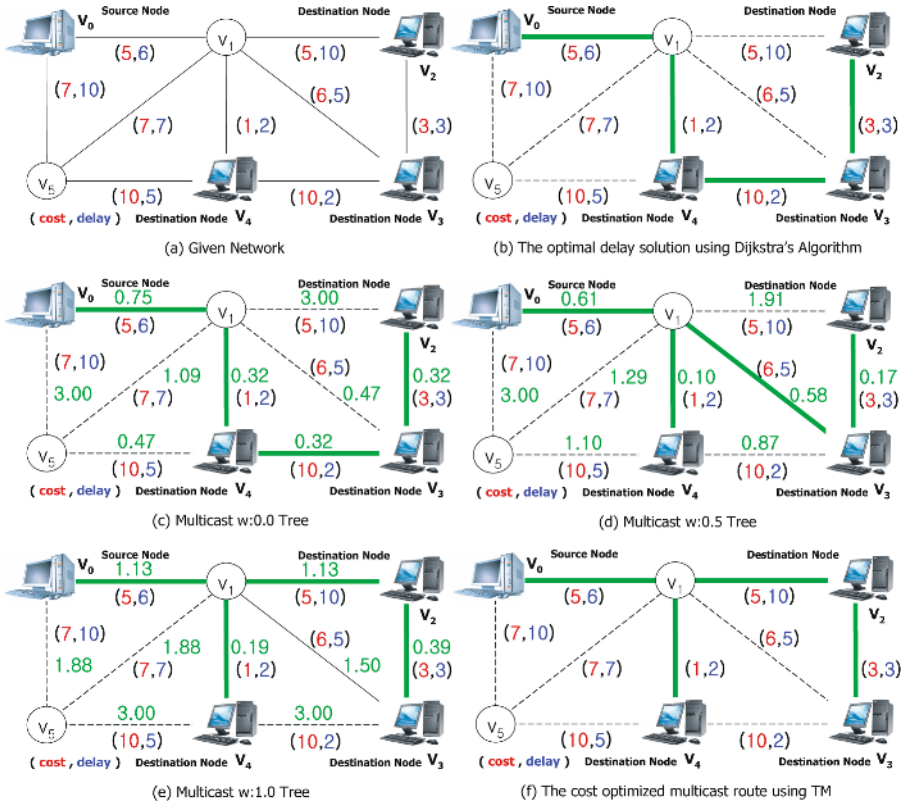


Fig. 3. Some algorithms, and the ω -Trees for each $\omega = 0.0, 0.5, \text{ and } 1.0$

We use the TM algorithm with N and construct the $T_{\omega:0.5}$ as Fig. 3 (d). Fig. 3 (c)-(e) show the trees constructed by the new parameter for each weight ω .

As indicated in Table 1, the tree cost sequence is $\phi_C(T_{TM}) \leq \phi_C(T_{\omega:1.0}) \leq \phi_C(T_{\omega:0.5}) \leq \phi_C(T_{\omega:0.0}) \leq \phi_C(T_{Dijkstra})$ and the tree delay sequence is $\phi_D(T_{Dijkstra}) \leq \phi_D(T_{\omega:0.0}) \leq \phi_D(T_{\omega:0.5}) \leq \phi_D(T_{\omega:1.0}) \leq \phi_D(T_{TM})$. Therefore, our method is quite likely performance of k^{th} shortest tree algorithm.

Table 1. The comparison with example results

$T_{Dijkstra}$ Fig. 3 (b)		$T_{\omega:0.0}$ Fig. 3 (c)		$T_{\omega:0.5}$ Fig. 3 (d)		$T_{\omega:1.0}$ Fig. 3 (e)		T_{TM} Fig. 3 (f)	
$\phi_C(T)$	$\phi_D(T)$	$\phi_C(T)$	$\phi_D(T)$	$\phi_C(T)$	$\phi_D(T)$	$\phi_C(T)$	$\phi_D(T)$	$\phi_C(T)$	$\phi_D(T)$
19	13	19	13	15	14	14	19	14	19

4 Performance Evaluation

4.1 Random Real Network Topology for the Simulation

Random graphs of the acknowledged model represent different kinds of networks, communication networks in particular. There are many algorithms and programs, but the speed is usually the main goal, not the statistical properties. In the last decade the problem was discussed, for example, by B. M. Waxman (1993) [19], M. Doar (1993, 1996) [5, 6], C.-K. Toh (1993) [18], E. W. Zegura, K. L. Calvert, and S. Bhattacharjee (1996) [20], K. L. Calvert, M. Doar, and M. Doar (1997) [3], R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal (2000) [12]. They have presented fast algorithms that allow the generation of random graphs with different properties, in particular, these are similar to real communication networks. However, none of them have discussed the stochastic properties of generated random graphs. A. S. Rodionov and H. Choo [16] have formulated two major demands for the generators of random graph: attainability of all graphs with required properties and uniformity of distribution. If the second demand is sometimes difficult to prove theoretically, it is possible to check the distribution statistically. The generation of random real network topologies is proposed by A. S. Rodionov and H. Choo [16], for the evaluation and the simulation results based on the network topology generated. The method uses parameter P_e , the probability of link existence between any node pair. We use the method by Rodionov and Choo.

4.2 Simulation Results

We now describe some numerical results, comparing the tree costs and the tree delays for each algorithms, respectively. Algorithms are the optimal delay solution with Dijkstra’s algorithm, the cost optimized multicast route with TM

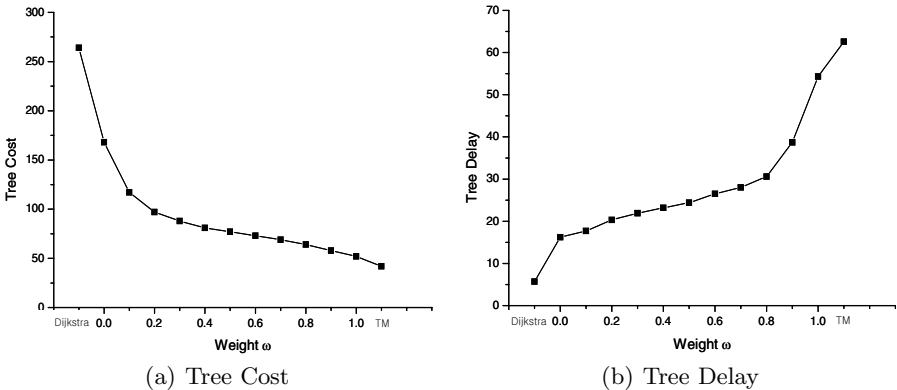


Fig. 4. $|M| : 15\%$ of $|V|=200$

algorithm, and ω -Trees. The proposed one is implemented in C . The 10 different network environments are generated for each size of given 200 nodes with $P_e=0.3$. A source s and destination nodes M , $|M| = 15\%$ of $|V| = 200$, are randomly selected in the network topology. We simulate 100 times (total $10 \times 100 = 1000$) for each network topology. Fig. 4 shows the simulation results for our method. The average tree cost is decreasing as ω is to be nearly 1. On the contrary if the average tree delay is increasing as ω is to be nearly 1. Therefore, ω plays on important role to combine two independent measures, the cost and the delay. If a delay bound is given, then we may find the tree which is appropriate for the minimum tree cost and acceptable the tree delay. Since the new parameter takes into both the cost and the delay consideration at the same time, it seems reasonable to use the new weight parameter.

5 Conclusion

We studied the efficient routing problem in the Steiner tree problem with QoS requirements. We formulated the new weight parameter which had taken into both the cost and the delay consideration at the same time. The cost of optimal delay solution is relatively more expensive than the cost of cost optimized multicast route, and moreover, the delay of the cost optimized multicast route is relatively higher than the delay of optimal delay solution. The weight ω plays on important role to combine two the measures. If the ω is nearly 0, then the tree delay is low. Otherwise the tree cost is low. Therefore if we decide the ω , then we find the efficient multicast routing tree. When network users have various QoS requirements, the proposed multicast weight parameter is very informative for them.

References

1. Y.-C. Bang and H. Choo, "On multicasting with minimum costs for the Internet topology," Springer-Verlag Lecture Notes in Computer Science, vol. 2400, pp.736-744, August 2002.
2. K. Bharath-Kumar and J. M. Jaffe, "Routing to multiple destinations in computer networks," IEEE Trans. Commun., vol. COMM-31, no. 3, pp. 343-351, March 1983.
3. K.L. Calvert, M. Doar, and M. Doar, "Modelling Internet Topology," IEEE Communications Magazine, pp. 160-163, June 1997.
4. E. Dijkstra, "A note on two problems in connexion with graphs," Numerische Mathematik, vol. 1, pp. 269-271, 1959.
5. M. Doar, Multicast in the ATM environment. PhD thesis, Cambridge Univ., Computer Lab., September 1993.
6. M. Doar, "A Better Mode for Generating Test Networks," IEEE Proc. GLOBE-COM'96, pp. 86-93, 1996.
7. M. R. Garey and D. S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W. H. Freeman and Co., San Francisco, 1979.
8. E. N. Gilbert and H. O. Pollak, "Steiner minimal tree," SIAM J. Appl. Math., vol. 16, 1968.

9. S. L. Hakimi, "Steiner's problem in graphs and its implication," *Networks*, vol. 1, pp. 113-133, 1971.
10. V. P. Kompella, J. C. Pasquale, and G. C. Polyzos, "Multicast routing for multimedia communication," *IEEE/ACM Trans. Networking*, vol. 1, no. 3, pp. 286-292, June 1993.
11. L. Kou, G. Markowsky, and L. Berman, "A fast algorithm for steiner trees," *Acta Informatica*, vol. 15, pp. 141-145, 1981.
12. R. Kumar, P. Raghavan, S. Rajagopalan, D Sivakumar, A. Tomkins, and E Upfal, "Stochastic models for the Web graph," *Proc. 41st Annual Symposium on Foundations of Computer Science*, pp. 57-65, 2000.
13. A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. McGraw-Hill, 2002.
14. R.C. Prim, "Shortest Connection Networks And Some Generalizations," *Bell System Techn. J.* 36, pp. 1389-1401, 1957.
15. S. Ramanathan, "Multicast tree generation in networks with asymmetric links," *IEEE/ACM Transactions on Networking*, vol. 4, no. 4, pp. 558-568, 1996.
16. A.S. Rodionov and H. Choo, "On Generating Random Network Structures: Connected Graphs," *Springer-Verlag Lecture Notes in Computer Science*, vol. 3090, pp. 483-491, September 2004.
17. H. Takahashi and A. Matsuyama, "An approximate solution for the steiner problem in graphs," *Mathematica Japonica*, vol. 24, no. 6, pp. 573-577, 1980.
18. C.-K. Toh, "Performance Evaluation of Crossover Switch Discovery Algorithms for Wireless ATM LANs," *IEEE Proc. INFOCOM'96*, pp. 1380-1387, 1993.
19. B. W. Waxman, "Routing of multipoint connections," *IEEE J-SAC*, vol. 6, no. 9, pp. 1617-1622, December 1988.
20. E.W. Zegura, K. L. Calvert, and S. Bhattacharjee, "How to model an Internet-network," *Proc. INFOVCOM'96*, pp. 594-602, 1996.
21. Q. Zhu, M. Parsa, and J. J. Garcia-Luna-Aceves, "A source-based algorithm for near-optimum delay-constrained multicasting," *Proc. IEEE INFOCOM'95*, pp. 377-385, March 1995.

The Gateways Location and Topology Assignment Problem in Hierarchical Wide Area Networks: Algorithms and Computational Results

Przemyslaw Ryba and Andrzej Kasprzak

Wroclaw University of Technology, Chair of Systems and Computer Networks,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
przemyslaw.ryba@pwr.wroc.pl
andrzej.kasprzak@pwr.wroc.pl

Abstract. This paper studies the problem of designing two-level hierarchical structure wide area network. The goal is to select gateways location, 2nd level network topology, channels capacities and flow routes in order to minimize the total average delay per packet in hierarchical network subject to budget constraint. The problem is NP-complete. Then, the branch and bound method is used to construct the exact algorithm. Also a heuristic algorithm is proposed. Some computational results are reported. Based on computational experiments, several properties of the considered problem, important from practical point of view, are formulated.

1 Introduction

Designing the huge wide area networks (WAN) containing hundreds of hosts and communication links is very difficult and demanding task. Design procedures (algorithms) suitable for small and moderate-sized networks, when applied directly to large networks, become very costly (from computational point of view) and sometimes infeasible. Design of huge wide area networks requires decomposition of design process. Common way to decompose design of huge WAN is to introduce hierarchical network topology [1]. In hierarchical WAN, nodes are grouped in clusters on the 1st level hierarchy, which in turn are grouped into 2nd level cluster and so on. The communication network of each cluster can be designed separately. In each cluster special communication nodes called “gateways”, which provide communication between nodes from different 1st level networks, are selected. Traffic between nodes in the same cluster uses paths restricted to local communication network. Traffic between nodes in different 1st level networks is first sent to local gateway, then via 2nd level network of gateways is sent to gateway located in destination 1st level network to finally reach the destination node. Example of structure of the hierarchical wide area network is presented in the Fig. 1. In the paper the exact and heuristic algorithms for simultaneous gateways location, network topology, channels capacity and flow assignment in two-level hierarchical wide area network are presented. The considered problem is formulated as follows:

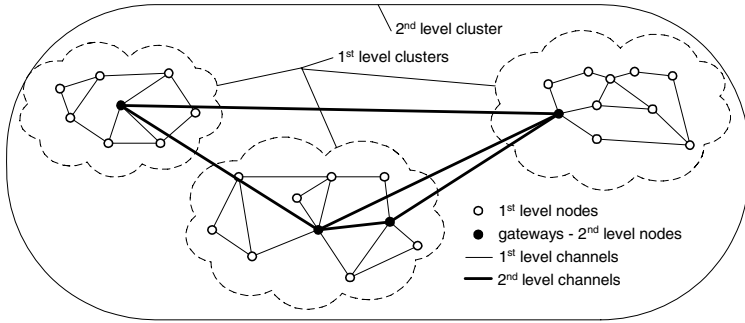


Fig. 1. Structure of the hierarchical wide area network

- given: topology of 1st level networks, potential gateways locations, set of potential 2nd level network channels and their possible capacities and costs (i.e. cost-capacity function), traffic requirements, budget of the network,
- minimize: total average delay per packet given by Kleinrock’s formula [2],
- over: gateways locations, 2nd level network topology, channel capacities, multicommodity flow (i.e. routing),
- subject to: multicommodity flow constraints, channel capacity constraints, budget constraint.

Discrete cost-capacity function considered here is most important from the practical point of view as channels capacities are chosen from the sequence defined by ITU-T recommendations. The problem formulated above is NP-complete. In section 3 of the paper NP-completeness of considered problem is proven.

Some algorithms for hierarchical network design can be found in [5], [7]. However, they are limited to tree topology of hierarchical network. In the paper [6] algorithm for router location, which includes simplified topology assignment without allocating capacities to channels, is presented. Also, in [5] algorithm for interconnecting two WANs is presented. In the paper [7] it was noticed that it is necessary to design network topology and gateway locations jointly. However, presented in [7] problem was limited to joint gateway location and capacity assignment in tree topology of the hierarchical network and heuristic algorithm was proposed only.

This paper joins problem of locating gateways in hierarchical wide area network with topology and capacity assignment problem, i.e. simultaneously gateways locations and topology of 2nd level network and capacities of channels of 2nd level network are selected. Thus, problem presented in the paper is more general and more important from practical point of view than problems considered in the literature.

2 Problem Formulation

Consider a hierarchical wide area network consisting of K networks on 1st level of hierarchy, each denoted by S_1^l , $l = 1, \dots, K$, and one 2nd level network denoted by S_2 . Let N_1^l be set of nodes and L_1^l set of channels of 1st level network S_1^l . Let n be the

total number of nodes in hierarchical network. Set of nodes N_2 consists of selected gateways and L_2 is set of channels connecting them. Let m be number of potential channels in 2nd level of hierarchical WAN and p – total number of channels. For each potential channel i there is the set $\bar{C}^i = \{c_1^i, \dots, c_{s(i)-1}^i\}$ of alternative values of capacities from which exactly one must be chosen if the channel was chosen to build the 2nd level network. Let d_k^i be the cost of leasing capacity c_k^i [\$/month]. Let $c_{s(i)}^i = 0$ for $i = 1, \dots, m$. Then $C^i = \bar{C}^i \cup \{c_{s(i)}^i\}$ be the set of alternative capacities from among which exactly one must be used to channel i . If the capacity $c_{s(i)}^i$ is chosen then the channel i is not used to build the network. Let x_k^i be the discrete variable for choosing one of available capacities for channel i defined as follows: $x_k^i = 1$, if the capacity c_k^i is assigned to channel i and $x_k^i = 0$, otherwise. Since exactly one capacity from the set C^i must be chosen for channel i , the following condition must be satisfied:

$$\sum_{k=1}^{s(i)} x_k^i = 1 \text{ for } i = 1, \dots, m \tag{1}$$

Let $X^i = \{x_1^i, \dots, x_{s(i)}^i\}$ be the set of variables x_k^i corresponding to the i -th channel. Let X_r' be the permutation of values of variables x_k^i , $i = 1, \dots, m$ satisfying the condition (1), and let X_r be the set of variables which are equal to 1 in X_r' .

Let denote by H^l set of gateways to place in network S_1^l and by J_g set of possible locations for gateway g . Let y_a^g be the discrete variable for choosing one of available locations for gateway g defined as follows: $y_a^g = 1$, if gateway g is located in node a and $y_a^g = 0$, otherwise. Each gateway must be placed in exactly one node; thus, it is required to satisfy following condition:

$$\sum_{a \in J_g} y_a^g = 1, \quad g \in H^l, \quad l = 1, \dots, K \tag{2}$$

Let Y_r' be the permutation of values of all variables y_a^g for which condition (2) is satisfied and let Y_r be the set of variables which are equal to 1 in Y_r' . The pair of sets (X_r, Y_r) is called a selection. Each selection (X_r, Y_r) determines locations of gateways and channels capacities in the 2nd level of hierarchical wide area network. Let \mathfrak{X} be the family of all selections.

Let r_{ij} be the average packet rate transmitted from node i to node j in the hierarchical network. It is assumed that $r_{ii} = 0$ for $i = 1, \dots, n$.

Let $T(X_r, Y_r)$ be the minimal average delay per packet in the hierarchical network, in which values of channels capacities are given by X_r and locations of gateways are

given by Y_r . $T(X_r, Y_r)$ can be obtained by solving a multicommodity flow problem in the network [2], [3]:

$$T(X_r, Y_r) = \min_{\underline{f}} \frac{1}{\gamma} \sum_{i \in L} \frac{f_i}{c^i - f_i} \tag{3}$$

subject to: \underline{f} is a multicommodity flow satisfying the requirements r_{ij} $i, j = 1, \dots, n$ and $f_i \leq c^i$ for every channel $i = 1, \dots, p$, where $\underline{f} = [f_1, \dots, f_p]$ is the vector of multi-commodity flow, f_i is the total average bit rate on channel i , and γ is the total packet arrival rate from external sources at all nodes of the wide area network. Then, the considered gateway location, topology assignment problem in hierarchical wide area network can be formulated as follows:

$$\min_{(X_r, Y_r)} T(X_r, Y_r) \tag{4}$$

subject to:

$$(X_r, Y_r) \in \mathfrak{R} \tag{5}$$

$$d(X_r, Y_r) = \sum_{x_k^i \in X_r} x_k^i d_k^i \leq B \tag{6}$$

where B denotes the budget of the hierarchical wide area network.

3 The Branch and Bound Algorithm

Assuming that $|H^l| = 1$ for $l = 1, \dots, K$, and $C^i = \bar{C}^i$ for $i = 1, \dots, p$, the problem (4-6) is resolved itself into the “topology, capacity and flow assignment problem” which is known as a NP-complete [3]. Since the problem (4-6) is more general it is also NP-complete. Then, the branch and bound method can be used to construct the exact algorithm for solving the considered problem. The detailed description of the calculation scheme of the branch and bound method may be found in the papers [8], [9].

The branch and bound method involves constructing two important operations specific for considered here problem: branching rules and lower bound. These operations are presented in following sections.

3.1 Branching Rules

The purposes of branching rules is to find the normal variable from the selection (X_r, Y_r) for complementing and generating a successor (X_s, Y_s) of the selection (X_r, Y_r) with the least possible value of the criterion function (3). We can choose a variable x_k^i or a variable y_a^g . For fixed variables y_a^g the problem (4-6) may be resolved into classical topology design problem. Then we can use the choice criterion Δ_{kj}^r on variables x_k^i presented in the paper [4].

To formulate the choice criterion on variables y_a^g we use the following theorem:

Theorem 1. Let $(X_r, Y_r) \in \mathfrak{R}$. If the selection (X_s, Y_s) is obtained from the selection (X_r, Y_r) by complementing the variable $y_a^g \in Y_r$ by the variable $y_b^g \in X_s$ then $T(X_s, Y_s) \leq T(X_r, Y_r) - \delta_{ab}^{gr}$, where

$$\delta_{ab}^{gr} = \begin{cases} \frac{1}{\gamma} \left(\sum_{i \in L} \frac{f_i}{c^i - f_i} - \sum_{i \in L_1^a} \frac{\tilde{f}_i}{c^i - \tilde{f}_i} + \sum_{i \in L-L_1^a} \frac{f_i}{c^i - f_i} \right) & \text{if } \tilde{f}^i < c^i \text{ for } i \in L_1^a \\ \infty & \text{otherwise} \end{cases} \tag{7}$$

$\tilde{f}_i = f_{ir} - f'_{ia} + f''_{ia}$: f''_{ia} corresponds to the packets flow between the nodes of network S_1^l and nodes of the remaining 1st level networks via the gateway located at node a , f'_{ia} corresponds to the packets exchanged between the nodes of network S_1^l and nodes of the remaining 1st level networks via the gateway located at node b after reallocating gateway from node a to node b .

Let $E_r = (X_r \cup Y_r) - F_r$, and let G_r be the set of all reverse variables of normal variables, which belong to the set E_r , where F_r is the set of variables constantly fixed in the r -th iteration of branch and bound algorithm. We want to choose a normal variable the complementing of which generates a successor with the possible least value of total average delay per packet. We should choose such pairs $\{(y_a^g, y_b^g): y_a^g \in E_r, y_b^g \in G_r\}$ or $\{(x_k^i, x_j^i): x_k^i \in E_r, x_j^i \in G_r\}$ for which the value of criterion δ_{ab}^{gr} or Δ_{kj}^{ir} is maximal.

3.2 Lower Bound

The lower bound LB_r of the criterion function (3) for every possible successor (X_s, Y_s) generated from the selection (X_r, Y_r) may be obtained by relaxing some constraints in the problem (4-6). To find the lower bound LB_r we reformulate the problem (4-6) in the following way:

- we assume that the variables x_k^i and y_a^g are continuous variables,
- we approximate the discrete cost-capacity curves (given by the set C^i) with the lower linear envelope [2]. Then, the constraint (6) may be relaxed by the constraint $\sum d^i c^i \leq B$, where $d^i = \min_{x_k^i \in X^i} (d_k^i / c_k^i)$ and c^i is the capacity of the channel i (continuous variable).

The solution of such reformulated problem can be found using the method proposed and described in the paper [10].

4 Heuristic Algorithm

The presented exact algorithm involves the initial selection $(X_1, Y_1) \in \mathfrak{R}$ for which the constraints (5) and (6) are satisfied [8]. Moreover, the initial selection should be the near-optimal solution of the problem (4-6). To find the initial selection the following heuristic algorithm is proposed. This heuristic algorithm may be also used to design of the hierarchical WAN when the optimal solution is not necessary.

- Step 1. For each gateway g , choose initial location from the set J_g such that cost of 2nd level network is minimal. Next, solve the classical topology assignment problem [4]. If this problem has no solution then the algorithm terminates – the problem (4-6) has no solution. Otherwise, perform $T^* = T$, where T is average packet delay in the network obtained by solving the topology assignment problem.
- Step 2. Choose the pair of gateway locations with maximal value δ_{ab}^{gr} given by expression (7). Change location of gateway on this indicated by that expression.
- Step 3. Solve the topology assignment problem. If the obtained value T is lower than T^* then perform $T^* = T$ and go to step 2. Otherwise, the algorithm terminates. The feasible solution is found. The network topology and gateways locations associated with the current T^* is heuristic solution of the problem (4-6).

5 Computational Results

The presented exact and heuristic algorithms were implemented in C++ code. Extensive numerical experiments have been performed with these algorithms for many different hierarchical network topologies and for many possible gateway locations. The experiments were conducted with two main purposes in mind: first, to examine the impact of various parameters on solutions to find properties of the problem (4-6) important from practical point of view and second, to test the computational efficiency of the algorithms.

The dependence of the optimal average delay per packet T on the budget B has been examined. In the Fig. 2 the typical dependence of T on the budget B is presented for different values of the average packet rates from external sources transmitted between each pair of nodes of the hierarchical network.

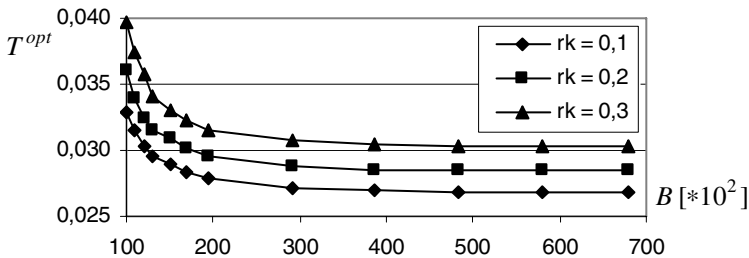


Fig. 2. The dependence of the criterion function T on the budget B

It follows from the Fig. 2 that there exists such budget B^* , that the problem (4-6) has the same solution for each B greater or equal to B^* . It means that the optimal solution of the problem (4-6) is on the budget constraint (6) for $B \leq B^*$ and it is inside the set of feasible solutions for $B > B^*$.

Conclusion 1. In the problem (4-6), for fixed average packets rate from external sources, there exists such value B^* of the budget, that for each $B \geq B^*$ we obtain the same optimal solution. It means that for $B \geq B^*$ the constraint (6) may be substituted by the constraint $d(X_r, Y_r) \leq B^*$

Moreover, it follows from results obtained for many considered networks, that the optimal delay per packet in terms of budget B is the function of the following class:

$$T = \frac{u_1}{B + u_2} + u_3 \tag{8}$$

where u_1 , u_2 and u_3 are constant coefficients. To find the function (8) for some hierarchical network it is enough to know only the values of u_1 , u_2 and u_3 , which can be computed by solving the identification problem, e.g. using the Least Squares Method.

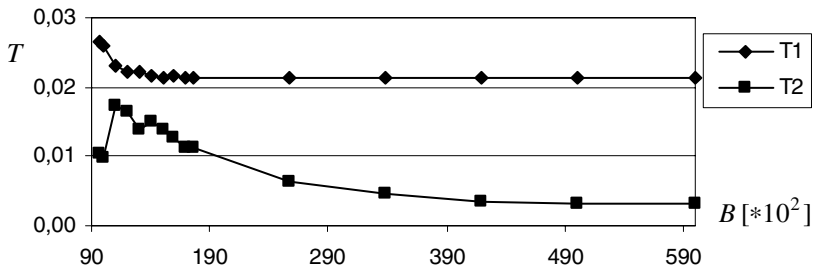


Fig. 3. Dependences of T_1 and T_2 on budget B

Let T_1 be the average delay per packet in any 1st level network and T_2 be the average delay per packet in 2nd level network obtained by solving the problem (4-6). The typical dependences of T_1 and T_2 on budget B are presented in the Fig. 3.

The observations following from the computer experiments may be formulated in the form of the conclusion below.

Conclusion 2. In the problem (4-6) there exists such value B' that for every $B > B'$, in each 1st level network, we obtain the same average delay per packet. Moreover, $B' < B^*$.

The observations presented as the conclusions 1 and 2 are very important from practical point of view. It shows that the influence of the investment cost (budget) on the optimal solution of the considered problem is limited, especially for greater values of budget B .

Let G be the number of gateways, which must be allocated in some 1st level network of the hierarchical network. The dependence of the optimal value of average

delay per packet T on number of gateways G has been examined. Fig. 4 shows the typical dependence of the delay T on number of gateways G .

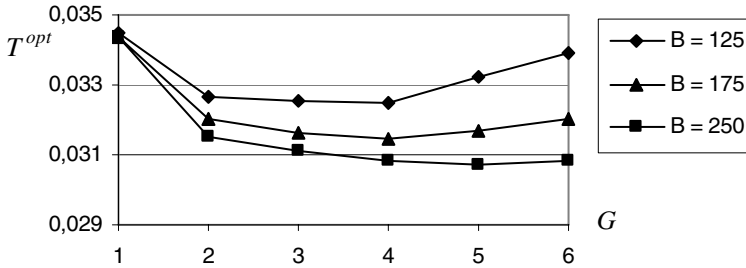


Fig. 4. The dependence of T^{opt} on number of gateways G

It follows from the numerical experiments and from the Fig. 4 that the function $T(G)$ is convex. Moreover, there exists the minimum of the function $T(G)$. The observations following from the computer experiments may be formulated in the form of the conclusions below.

Conclusion 3. The performance quality of the hierarchical wide area networks depends on the number of gateways in the 1st level networks.

Conclusion 4. There exists such value of the number of gateways G , in some 1st level network of the hierarchical network, for which the function $T(G)$ is minimal.

Let G_{min} be the number of gateways in some 1st level network such that $T(G_{min}) = \min T(G)$.

Conclusion 5. The number of gateways G_{min} in the hierarchical network, for which the value of the function $T(G)$ is minimal, depends on the budget B .

Because the algorithm for the gateways location and topology assignment problem assumes that the numbers of gateway for each 1st level network are given, then the properties described in conclusion 4 and 5 allow to simplify the design process of the hierarchical network. First, we may calculate the number of gateways for each 1st level network and next, we may solve the gateways location and topology assignment problem using the presented exact or heuristic algorithm.

In order to evaluate the computational efficiency of the exact algorithm many networks were considered. For each network the number of iteration of the algorithm was recorded and compared for all considered networks.

Let D_{max} be the maximal building cost of the network, and let D_{min} be the minimal building cost of the network; the problem (4-6) has no solution for $B < D_{min}$. To compare the results obtained for different hierarchical wide area networks topologies we introduce the normalized budget $\bar{B} = ((B - D_{min}) / (D_{max} - D_{min})) \cdot 100\%$.

Moreover, let $\Phi^i(\bar{B})$ be the number of iterations of the branch and bound algorithm to obtain the optimal value of T for normalized budget equal to \bar{B} for i -th considered network topology. Let

$$\Phi(u, v) = \frac{1}{Z} \sum_{i=1}^Z \left(\frac{\sum_{\bar{B} \in [u, v]} \Phi^i(\bar{B})}{\sum_{\bar{B} \in [0, 100]} \Phi^i(\bar{B})} \right) \cdot 100\%$$

be the arithmetic mean of the relative number of iterations for $\bar{B} \in [u, v]$ calculated for all considered network topologies and for different number of gateways, where Z is the number of considered wide area networks. Fig. 5 shows the dependency of Φ on divisions $[0\%, 10\%)$, $[10\%, 20\%)$, ..., $[90\%, 100\%]$ of normalized budget \bar{B} . It follows from Fig. 4 that the exact algorithm is especially effective from computational point of view for $\bar{B} \in [30\%, 100\%]$.

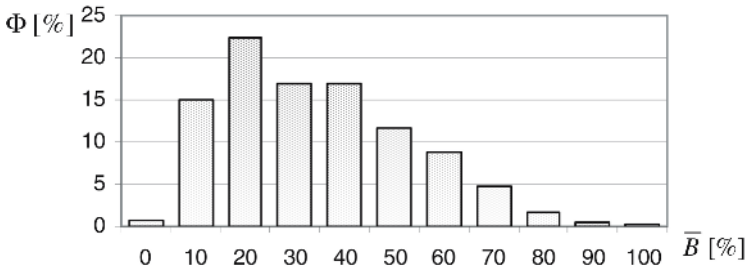


Fig. 5. The dependence of Φ on normalized budget \bar{B}

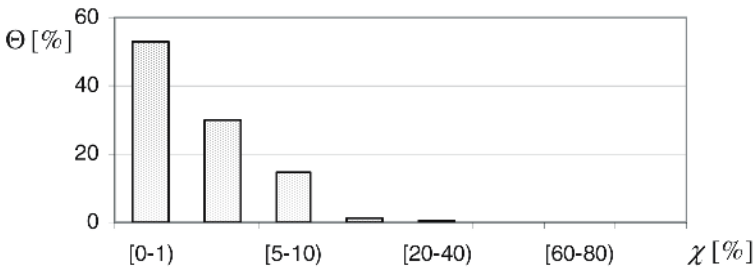


Fig. 6. The difference between heuristic and optimal solutions

The distance between heuristic and optimal solution has been also examined. Let T^{heur} be the solution obtained by heuristic algorithm and T^{opt} be the optimal value obtained by the exact algorithm for the problem (4-6). Let χ denote the distance between heuristic and optimal solutions: $\chi = \left| T^{heur} - T^{opt} \right| / T^{opt} \cdot 100\%$. The value χ shows how the results obtained using the heuristic algorithm are worse than the optimal solution. Let

$$\Theta[a, b] = \frac{\text{number of solutions for which } \chi \in [a, b]}{\text{number of all solutions}} \cdot 100\%$$

denote number of solutions obtained from heuristic algorithm (in percentage) which are greater than optimal solutions more than $a\%$ and less than $b\%$. Fig. 6 shows the dependence Θ on divisions [0%–1%), [1%–5%), ..., [80%–100%).

6 Conclusion

The exact and heuristic algorithms for solving the gateway location and network topology assignment problem in hierarchical network are presented. The considered problem is more general than the similar problems presented in the literature. It follows from computational experiments (Fig. 6) that more than 50% approximate solutions differ from optimal solutions at most 1%. It is necessary to stress that over 90% approximate solutions differ from optimal solutions at most 10%. Moreover, we are of opinion that the hierarchical network properties formulated as conclusions 4 and 5 are important from practical point of view. They say that there is the influence of the gateways number on performance quality of the hierarchical network.

Acknowledgment. This work was supported by a research project of The Polish State Committee for Scientific Research in 2005-2007.

References

1. Kleinrock L., Kamoun F.: Optimal Clustering Structures for Hierarchical Topological Network Design of Large Computer Networks. *Networks* 10 (1980) 221-248
2. Fratta, L., Gerla, M., Kleinrock, L.: The Flow Deviation Method: an Approach to Store-and-Forward Communication Network Design. *Networks* 3 (1973) 97-133
3. Kasprzak, A.: Topological Design of the Wide Area Networks. Wroclaw University of Technology Press, Wroclaw (2001)
4. Markowski M., Kasprzak A.: The Web Replica Allocation and Topology Assignment Problem in Wide Area Networks: Algorithms and Computational Results. *Lecture Notes in Computer Science* 3483 (2005) 772-781
5. Liang S.C., Yee J.R.: Locating Internet Gateways to Minimize Nonlinear Congestion Costs. *IEEE Transactions On Communications* 42, (1994) 2740-50
6. Liu Z., Gu Y., Medhi D.: On Optimal Location of Switches/Routers and Interconnection. Technical Report, University of Missouri-Kansas City (1998)
7. Saha D. and Mukherjee A.: On the multidensity gateway location problem for multilevel high speed network. *Computer Communications* 20 (1997) 576-585
8. Wolsey, L.A.: *Integer Programming*. Wiley-Interscience, New York (1998)
9. Walkowiak K.: A Branch and Bound Algorithm for Primary Routes Assignment in Survivable Connection Oriented Networks. *Computational Optimization and Applications* 27, Kluwer Academic Publishers (2004) 149-171
10. Markowski M., Kasprzak A.: An exact algorithm for host allocation, capacity and flow assignment problem in WAN. *Internet Technologies. Applications and Societal Impact*, Kluwer Academic Publishers, Boston (2002) 73-82

Developing an Intelligent Supplier Chain System Collaborating with Customer Relationship Management

Gye Hang Hong¹ and Sung Ho Ha²

¹ Hyundai Information Technology Research Institute, Yongin-Si,
Gyeonggi-do, 449-910, Korea

² School of Business Administration, Kyungpook National University,
Daegu, 702-701, Korea
hsh@mail.knu.ac.kr

Abstract. We propose an intelligent supply chain system collaborating with customer relationship management system in order to assess change in a supply partner's capability over a period of time. The system embeds machine learning methods and is designed to evaluate a partner's supply capability that can change over time and to satisfy different procurement conditions across time periods. We apply the system to the procurement and management of the agricultural industry.

1 Introduction

Nowadays the business environment is changing from the product-centered stage to customer-centered stage. Supply capacity of companies exceeds customer demands because of continuous improvement of production technologies. In addition, the globalization of trade and prevalence of communication networks make companies more compete with competitors. As a result, producers are not a market leader any more. Companies can not survive from the global competition without understanding customers. They must understand who their customers are, what they buy and when they buy, and even predict their behavior.

Customer-centered markets convert the traditional production environment to new one which requires quick responses to changes in customer needs. Companies have integrated various internal resources located inside the organization and constructed electric document interchange systems to exchange production information with their partners. They begin to invite production partners into their projects or businesses, organize a valuable supply chain network, and collaborate with them in many business areas. Therefore, the success of business depends on not competition among companies but competition among supply chain networks.

In this business environment, constructing a valuable supply chain determines the competitive power of companies: Quick response to customer needs, and effective selection and management of good supply partners. Therefore, we develop a decision support system for constructing an intelligent supply chain collaborating with the customer relationship management. In this system, we explain how manufacturers identify their valuable customers and how they establish competitive supply chains to select and manage their good supply partners.

2 Current Issues in Supply Chain Management

2.1 Evaluating Changes in Supply Capability Conditions over Time

Supply capability conditions of partners can change over time. It is difficult for supply partners to maintain the same capability conditions for all supply periods because of changes in delivery conditions, inventory levels, and the overall market environment [10]. In particular, it is hard when suppliers are selected for products which have seasonal demands and capability can fluctuate over time (e.g., agricultural products). Customer consumption patterns also vary over time in industries producing seasonal products. Therefore manufacturers must adopt different procurement strategies and evaluate their suppliers according to the changing consumption.

However, most supplier selection models did not consider that supply capabilities and procurement conditions could change over time. Recent studies have considered only comprehensive supply capability conditions during the total analyzed periods [2],[6],[8],[11],[12]. In this case, suppliers may not be selected due to the impact of the low quality of the last season, regardless of the fact that they were in good standing for this period. After a supplier has been chosen, a manufacturer is not able to take necessary action even though product quality could decrease after a certain period of time. Thus, it is important that a supply chain system divide all analyzed periods into meaningful period units (PFA), evaluate the supply conditions of each unit, and compile the results.

2.2 Evaluating Supply Partners with Multiple Criteria

A supply chain system has to consider multi-criteria, including quantitative and qualitative aspects, in order to evaluate a supply partner's capability. In an early study on supplier selection criteria, Dickson [3] identified 23 criteria considered by purchasing managers when selecting various partners. Since the Dickson study, many researchers have identified important criteria that have varied according to industry and purchasing situations. In his portfolio approach, Kraljic identified purchasing in terms of profit impact and supply risk [7]. Profit impact includes such elements as the expected monetary volume, which is tied with goods or services to be purchased and the impact on future product quality. Indicators of supply risk may include the availability of goods or services under consideration and the number of potential partners. Therefore, in this study, we choose such criteria as price, delivery, quality, and order quantity to evaluate the profit impact. In determining supply risk, we use the criteria of reputation and position, warranty and claim strategies, and the share of information.

2.3 Selecting Partners Who Satisfy the Procurement Conditions

Supply partners may be able to meet some, but not all of the procurement conditions. One partner may be able to meet price and quality demands, but it may not be able to satisfy quantity or delivery conditions. Another partner can guarantee quantity and delivery, but not price requirement. Due to different conditions among partners, manufacturers want to select supply partners that can satisfy all their needs as well as maximize their revenue. Optimal partners must be selected who not only maximize the revenue of a purchaser, but also manage a level of supply risk. Supply risk changes over a period of time. In order to appropriately manage the level of risk, it needs to be

measured by period. Then, partners can be qualified, either rigorously or less, according to the level of risk. A supply chain system has to qualify partners rigorously in the case of high risk and qualify partners less rigorously for the case of low risk.

3 Intelligent Supply Chain System (ISCS)

The ISCS consists of five main functions, as shown in Fig. 1. They include defining multiple criteria, dividing the total contract period into periods for assessment (PFA), establishing ideal purchasing conditions, segmenting supply partners by PFA, and selecting the optimal partners by PFA.

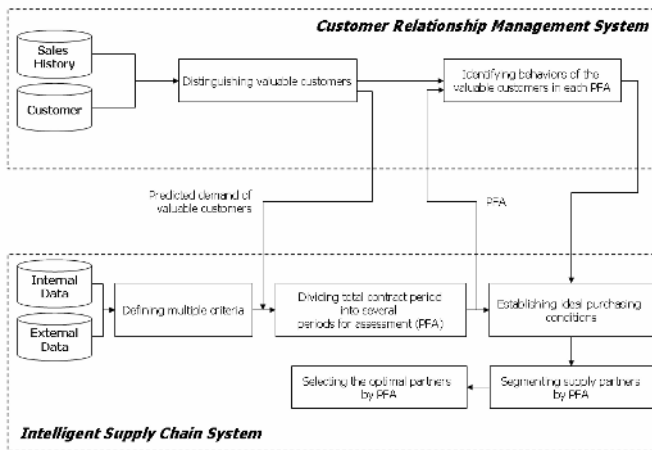


Fig. 1. An intelligent supply chain system collaborating with a customer relationship management system

The CRM system conjoins customer profiles and sales history, measures customer lifetime values or RFM (Recency, Frequency, and Monetary) values, and distinguishes valuable customers. The system informs the ISCS of valuable customers and their predicted demand. On the other hand, the ISCS defines important criteria with regard to assessing supply partners. The ISCS, then, divides the overall purchasing-contract period into PFAs from the viewpoint of supply market risk. The level of supply risk is defined as the difference (gap) between supply and demand for each period. The derived PFAs feed back to the CRM system.

The CRM system identifies buying behaviors of the valuable customers in each PFA. The system finds out important factors which customers consider when they make a purchase an item in each PFA. Customers may consider that quality is an important factor in a period; however, they can change their mind and think that price is more important in other periods. Therefore, a manufacturer must know important factors which have much effect on the buying behaviors of the customers. In addition, the manufacturer should know how much the change in customer satisfaction is sensitive to the change in values of the important factors.

Then, the ISCS searches the available supply partners and qualifies them in order to only select a small number of partners who have similar supply capability. In doing so, the ISCS segments all available partners into several groups having similar characteristics, and evaluates each group to choose qualified ones. The chosen partners have the same characteristics of supply capability, since the ISCS selects them from the same qualified groups.

To sum up, the CRM system is helpful in predicting customer demand, identifying behaviors of the valuable customers in each PFA, and discovering the purchasing factors of importance. With this useful information in hand, the ISCS can track changes in supply capability over time, select optimal supply partners whenever appropriate.

4 Application of ISCS to Seasonal Products

4.1 Distinguishing Valuable Customers

The CRM system identifies valuable customers in terms of RFM. It extracts RFM values from sales data and divides all the customers into several customer segments which have a similar RFM patterns. The RFM clustering is one of the methods for discovering customer segments [1], [4]. RFM is defined as follows: Recency is the time period of last purchase during the analyzed time period; Frequency is the number of purchases during the analyzed time period; Monetary is the amount of spent money during the analyzed time period.

The CRM system uses a self-organizing map (SOM) with three input nodes and three by three output nodes. Each input (i.e., recency, frequency, and monetary values) can fall into one of the following categories: above the average or below the average. Table 1 summarizes the segmentation results.

Table 1. Clustering customers into four segments in terms of RFM. The sign \uparrow means above the average and \downarrow means below the average.

Segment	R(ecency)	F(requency)	M(onetary)	No. of customers
1	R \uparrow (0.66)	F \uparrow (0.42)	M \uparrow (0.54)	20
2	R \downarrow (0.38)	F \uparrow (0.88)	M \uparrow (0.72)	4
3	R \downarrow (0.22)	F \downarrow (0.17)	M \downarrow (0.08)	46
4	R \uparrow (0.76)	F \downarrow (0.09)	M \downarrow (0.05)	70
Average	0.51	0.39	0.35	

As shown in Table 1, segment 1 is better than others in terms of RFM. Segment 2 is also superior to others in terms of FM. It implies that segments 1 and 2 contain valuable customers. However, segments 3 and 4 are less valuable groups, since they make a purchase small quantity of items during a short period.

4.2 Diving the Total Contract Period into Periods for Assessment

In order to assess the capability of a supply partner dynamically, the ISCS divides all periods into periods for assessment (PFA), in which the risk of the supply market is similar. A large difference between supply and demand indicates a low-market risk

because the manufacturer can find alternatives easily and is able to pay a low switching cost, when a partner does not deliver a product or service. A small difference between supply and demand entails a high-market risk.

The ISCS measures the market risk by period and then groups similar market-risk periods through a genetic algorithm and a linear regression model (See Eq. 1).

$$Fitness\ function = \alpha \sum_{i=1}^N R_i^2 w_i + \beta F(N) \tag{1}$$

where $R_i^2 = \max\{R_{i,exp}^2, R_{i,linear}^2, R_{i,log}^2\}$,

$$w_i = \frac{\text{no. of periods in the } i\text{th interval}}{\text{no. of whole periods}},$$

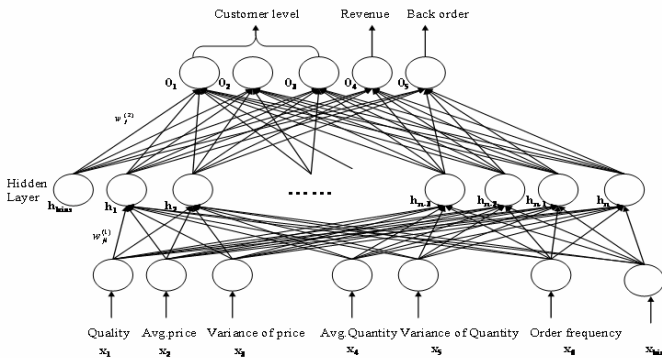
$N = \text{no. of intervals}$,

$R_i^2 = R^2$ (residual error) of the i th interval, $F(N) \propto 1/n$, $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$, $\alpha + \beta = 1$

When setting the total analyzed period to a single year, the ISCS obtains four meaningful periods for assessment.

4.3 Identifying Behaviors of the Valuable Customers in Each PFA

After the ISCS divides the overall period into PFAs, the CRM system identifies the important factors which have much influence on buying behavior patterns of valuable customers in each PFA. In doing so, it employs a neural network technique (see Fig. 2) to discover two kinds of important knowledge: *IF* (Important Factor) and *DCR* (Degree of Contribution/Risk) of each IF.



1. Measuring degree of importance of each factors

$$w_i = \frac{(\sum_L \frac{|p^0 - p^i|}{p^0})}{n}$$

2. Measuring degree of contribution/risk of each factors

$$R_i = \sum_{j=1}^6 ((w_j^{(1)})^2 \times (w_j^{(2)})^2 \times \text{var}(\text{sigm}(\sum_{k=1}^6 w_k^{(1)} \times x_k)))$$

Fig. 2. A neural network model for discovering the knowledge regarding customer buying behaviors

The NN has five output nodes: three nodes to classify the customer level, one node to predict the revenue, and one node to predict the number of back orders. In order to represent the customer level, the customer segment number is used. That is, (1, 0, 0) denotes segment 1, (0, 1, 0) denotes segment 2, (0, 0, 1) represents segments 3 and 4. The NN has one hidden layer, six input nodes, and one bias. Six inputs include quality, average price, variance of price, average quantity, variance of quantity, and order frequency.

To discover IF, the NN uses a sensitivity method, a feature weighting method [9]. The sensitivity of input node is calculated by removing the input node from the trained neural network (See Eq. 2).

$$w_i = \frac{(\sum_L |p^0 - p^i|)}{n} \tag{2}$$

where p^0 is the normal prediction value for each training instance after training, and p^i is the modified prediction value when the input node i is removed. L is the set of training data, and n is the number of training data.

The NN also uses an activity method to measure the DCR of each IF. The activity method measures the variance of activation level of an input node (see Eq. 3).

$$R_i = \sum_{j=1}^n ((w_{ji}^{(1)})^2 \times (w_j^{(2)})^2 \times \text{var}(\text{sigm}(\sum_{i=1}^6 w_{ji}^{(1)} \times x_i))) \tag{3}$$

Table 2 summarizes several IFs and their DCRs by each PFA.

Table 2. Derived key factors of valuable customers over periods in time. Note that P denotes Prediction and C for Classification.

PFA		t1		t2		t3		t4	
		P	C	P	C	P	C	P	C
Quality	Level	1 st – 2 nd level		4 th – 7 th level		1 st – 2 nd level		1 st level	
	IF	0.652	0.176	0.473	0.085	0.345	0.042	0.486	0.112
	DCR	0.329		0.426		0.289		0.019	
Fre- quency	Level	High		Average		Average		Above average	
	IF	0.072	0.206	0.024	0.026	0.002	0.116	0.574	0.418
	DCR	0.010		0.042		0.042		0.356	
Price	Level	Above average		Average		Above average		High	
	IF	0.432	0.176	0.517	0.341	0.365	0.304	0.045	0.074
	DCR	0.044		0.506		0.427		0.001	
Quan- tity	Level	High		Average		Average		Below average	
	IF	0.334	0.059	0.023	0.286	0.195	0.019	0.462	0.007
	DCR	0.254		0.001		0.174		0.316	
Derived key factors		Quality, Quantity		Quality, Price		Quality, Price		Frequency, Quantity	

Fig. 3 illustrates correlation relationships between IF and DCR in the different PFA. The larger values both IF and DCR have (upper right position in each graph), the more important IF is.

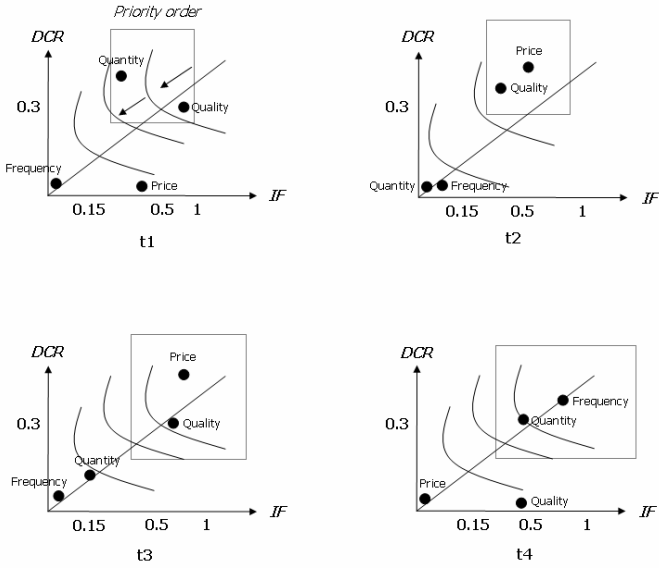


Fig. 3. Correlation relationships between IF and DCR in the different PFA

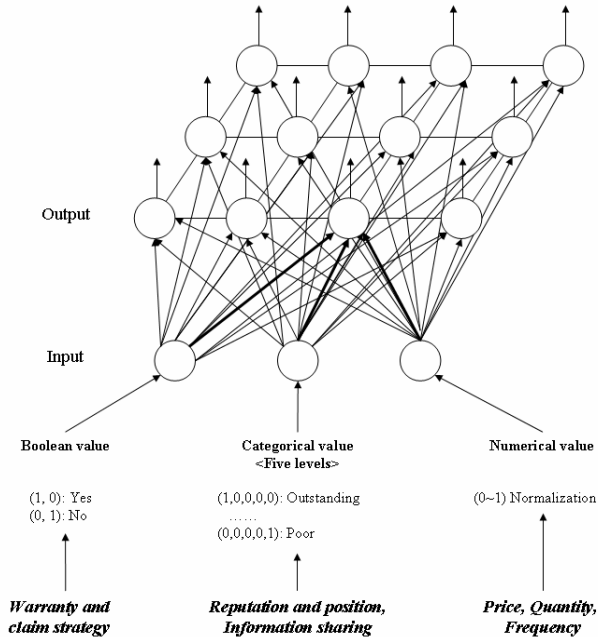


Fig. 4. The SOM network using various types of input data. The learning function is $w_j(n+1) = w_j(n) + l_r(n)h_{j,i(x)}(n)[x(n) - w_j(n)]$ where the numerical value is $(x - w_j)$, and the Boolean or categorical values are 1, $x = w_j$, otherwise, 0, $x \neq w_j$.

4.4 Segmenting Supply Partners by PFA

The ISCS segments supply partners into several groups which have similar supply conditions by using a SOM network—a special neural network [5]. Since the criteria for assessing partners consist of both quantitative and qualitative variables, the input data types for a SOM are different. Warranty and claim strategy criterion uses Boolean data, ‘yes (1)’ or ‘no (0)’. Reputation and position, and information sharing criteria use categorical data that include the following scale: outstanding (5), very good (4), good (3), average (2), and poor (1). The other criteria use numerical data. Therefore, in order to use the SOM, each data type is appropriately modified. Fig. 4 shows a SOM network with input variables.

4.5 Selecting the Optimal Supply Partners

After clustering the supply partners, the ISCS evaluates the supply conditions of the supplier groups and qualifies them as follows:

1. In the case of a high-supply risk, among the supplier groups which have a higher-than-average value with regard to all qualitative criteria (quality, reputation and position, warranty and claim strategy, and sharing of information), the ISCS selects the candidate groups which are closest to the ideal purchasing conditions;
2. In the case of a low-supply risk, among all supplier groups, the ISCS selects the candidate groups which are nearest to the ideal purchasing conditions;
3. The ISCS conducts this process for all PFAs and obtains the qualified partner groups and qualified suppliers.

Table 3 shows the supply conditions of the qualified groups in each PFA.

Table 3. Supply conditions of the qualified groups. Note that QN stands for quantity, F for frequency, P for price, QL for quality, R for reputation and position, W for warranty and claim, and I for information sharing.

PFA	Group	QN	F	P	QL	R	W	I	# of partners
<i>t1</i>	1	0.34	0.36	0.52	0.98	4	1	3	11
	4	0.37	0.38	0.42	0.87	3	1	3	3
	7	0.45	0.14	0.53	0.97	5	1	4	6
	Avg	0.35	0.16	0.62	0.74	-	-	-	47
<i>t2</i>	2	0.16	0.09	0.32	0.80	4	1	3	4
	4	0.38	0.31	0.48	0.97	5	1	4	15
	5	0.22	0.32	0.31	0.75	2	0	3	4
	Avg	0.30	0.10	0.52	0.76	-	-	-	98
<i>t3</i>	2	0.45	0.46	0.68	0.97	5	1	4	6
	4	0.31	0.23	0.32	0.80	4	1	3	1
	5	0.29	0.25	0.19	1.0	5	1	4	3
	Avg	0.22	0.15	0.41	0.79	-	-	-	30
<i>t4</i>	2	0.35	0.35	0.59	0.97	5	1	4	17
	5	0.20	0.28	0.42	0.89	5	1	4	11
	Avg	0.12	0.11	0.28	0.67	-	-	-	89

5 Conclusion

In this paper, we suggested an intelligent supply chain system collaborating with a customer relationship management system. The intelligent SCS adopted the following methods to solve the current issues on supply chains:

1. The ISCS divided the total analyzed period into meaningful PFAs, segmented suppliers into several clusters which have similar supply conditions within each PFA, and evaluated the clusters.
2. The system handled with the multiple criteria of quantitative and qualitative types with regard to a partner's risks and profits. Then, the system chose supply partners who were deemed to be the most valuable from the selected clusters. It was able to solve the complexity of the problem by using those criteria.

We applied this solution to the agricultural industry and illustrated the results.

References

1. Bult, J.R., Wansbeek, T.J.: Optimal Selection for Direct Mail. *Marketing Science* 14 (1995) 378–394
2. De Boer, L., Labro, E., Morlacchi, P.: A Review of Methods Supporting Supplier Selection. *European Journal of Purchasing and Supply Management* 7 (2001) 75–89
3. Dickson, G.W.: An Analysis of Vendor Selection Systems and Decisions. *Journal of Purchasing* 2 (1966) 5–17
4. Ha, S.H., Park, S.C.: Application of Data Mining Tools to Hotel Data Mart on the Intranet for Database Marketing. *Expert System with Applications* 15 (1998) 1–31
5. Han, J., Kamber, M.: *Data Mining—Concepts and Techniques*. Morgan Kaufmann, CA (2001)
6. Holt, G.D.: Which Contractor Selection Methodology?. *International Journal of Project Management* 16 (1998) 153–164
7. Kraljic, P.: Purchasing Must Become Supply Management. *Harvard Business Review* 61 (1983) 109–117
8. Lee, E.K., Ha, S., Kim, S.K.: Supplier Selection and Management System Considering Relationships in Supply Chain Management. *IEEE Transactions on Engineering Management* 48 (2001) 307–318
9. Shin, C.K., Yun, U.T., Kim, H.K., Park, S.C.: A Hybrid Approach of Neural Network and Memory-Based Learning to Data Mining. *IEEE Trans. on Neural Networks* 11 (2000) 637–646
10. Talluri, S., Sarkis, J.: A Model for Performance Monitoring of Suppliers. *International Journal of Production Research* 40 (2002) 4257–4269
11. Weber, C.A., Current, J.R., Desai, A.: Non-cooperative Negotiation Strategies for Vendor Selection. *European Journal of Operational Research* 108 (1998) 208–223
12. Weber, C.A., Desai, A.: Determination of Path to Vendor Market Efficiency Using Parallel Coordinates Representation: a Negotiation Tool for Buyers. *European Journal of Operational Research* 90 (1996) 142–155

The Three-Criteria Servers Replication and Topology Assignment Problem in Wide Area Networks

Marcin Markowski and Andrzej Kasprzak

Wroclaw University of Technology, Chair of Systems and Computer Networks,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
marcin.markowski@pwr.wroc.pl,
andrzej.kasprzak@pwr.wroc.pl

Abstract. The designing of wide area networks is usually an optimization process with accurately selected optimization criterion. The most utilized criterions are the quality of service in the network (indicated by average packet delay), the capacity cost (channel capacity leasing cost) and server costs (cost of connecting servers or replicas at nodes). This paper studies the problem of designing wide area networks with taking into account those three criteria. Then, the goal is select servers replica allocation at nodes, network topology, channel capacities and flow routes in order to minimize the linear combination of average delay per packet, capacity cost and server cost. The problem is NP-complete. An exact algorithm, based on the branch and bound method is proposed. Some computational results are reported and several properties of the considered problem are formulated.

1 Introduction

During the Wide Area Network (WAN) designing process, different optimization criterions may be taken into account. In the literature the one-criteria problems are often studied [1]. The very often criterion is the quality of service in the network [2] and the different kind of investment costs. There are also some algorithms for solving two-criteria problems, proposed in [3]. In fact, there are two basic kinds of the investment costs in the wide area networks. First of them, the network support cost must be borne regularly, for example once a month or once a year. The typical example is the capacity cost (leasing cost of channels). Second kind is disposable cost, borne once when the network is built or expanded. The example of disposable cost is server cost (connecting cost of servers at nodes). In the literature there are no solutions with two different cost criteria. Such approach may be very useful, because WAN designers often can not compare server cost with capacity cost. Then we consider the three-criteria design problem in wide area networks, where the linear combination of the quality of service, capacity cost and server cost are incorporated as the optimization criterion.

Designing of network consists of topology, capacity and flow assignment and resource allocation. Resources often have to be replicated in order to satisfy all users demands [4], [5]. In the paper an exact algorithm for simultaneous assignment of

server's replica allocation, network topology, channels capacity and flow assignment is proposed. We use the combined criterion composed of the three indicators, representing quality of the network and different network costs. Moreover, we denote, that the capacity cost is limited. The considered problem is formulated as follows:

- given: user allocation at nodes, for every server the set of nodes to which server may be connected, number of replicas, budget of the network, traffic requirements user-user and user-server, set of potential channels and their possible capacities and costs (i.e. cost-capacity function)
- minimize: linear combination of the total average delay per packet, the capacity cost and the server cost
- over: network topology, channel capacities, multicommodity flow (i.e. routing), replica allocation
- subject to: multicommodity flow constraints, channel capacity constraints, budget (maximal capacity cost) constraint, replica allocation constraints.

We assume that channels' capacities can be chosen from discrete sequence defined by international ITU-T recommendations. Then, the formulated above three-criteria servers replication and topology assignment problem is NP-complete [1], [6].

Some solutions for the wide area network optimization problems can be found in the literature. The topology assignment problem and capacity and flow assignment problem are considered in papers [1], [6], [7] and resource allocation and replication in [3], [4], [5]. In the paper [3] the two-criteria web replica allocation and topology assignment problem is considered. Problem presented in this paper is much more general than problem presented in [3], because we consider three criteria: two criteria the same as in the paper [3] and a new one – the capacity cost as the third criterion. Then, the three-criteria problem is more complicated and much more useful for wide area network designers. Such problem, with different costs optimization criteria have been not considered in the literature yet.

2 Problem Formulation

Consider a WAN consisting of N nodes and b potential channels which may be used to build the network. For each potential channel i there is the set $\bar{C}^i = \{c_1^i, \dots, c_{s(i)-1}^i\}$ of alternative values of capacities from which exactly one must be chosen if the i -th channel was chosen to build the WAN. Let d_j^i be the cost of leasing capacity c_j^i [€/month]. Let $c_{s(i)}^i = 0$ for $i = 1, \dots, b$. Then $C^i = \bar{C}^i \cup \{c_{s(i)}^i\}$ be the set of alternative capacities from among which exactly one must be used to channel i . If the capacity $c_{s(i)}^i$ is chosen then the i -th channel is not used to build the wide area network. Let x_j^i be the decision variable, which is equal to one if the capacity c_j^i is assigned to channel i and x_j^i is equal to zero otherwise. Since exactly

one capacity from the set C^i must be chosen for channel i , then the following condition must be satisfied:

$$\sum_{j=1}^{s(i)} x_j^i = 1 \text{ for } i = 1, \dots, b. \tag{1}$$

Let $W^i = \{x_1^i, \dots, x_{s(i)}^i\}$ be the set of variables x_j^i , which correspond to the i -th channel. Let X_r' be the permutation of values of all variables x_j^i for which the condition (1) is satisfied, and let X_r be the set of variables, which are equal to one in X_r' .

Let K denotes the total number of servers, which must be allocated in WAN and let LK_k denotes the number of replicas of k -th server. Let M_k be the set of nodes to which k -th server (or replica of k -th server) may be connected, and let $e(k)$ be the number of all possible allocation for k -th server. Since only one replica of server may be allocated in one node then the following condition must be satisfied

$$LK_k \leq e(k) \text{ for } k = 1, \dots, K. \tag{2}$$

Let y_{kh} be the decision binary variable for k -th server allocation; y_{kh} is equal to one if the replica of k -th server is connected to node h , and equal to zero otherwise. Since LK_k replicas of k -th server must be allocated in the network then the following condition must be satisfied

$$\sum_{h \in M_k} y_{kh} = LK_k \text{ for } k = 1, \dots, K. \tag{3}$$

Let Y_r be the set of all variables y_{kh} , which are equal to one. The pair of sets (X_r, Y_r) is called a selection. Let \mathfrak{R} be the family of all selections. X_r determines the network topology and capacities of channels and Y_r determines the replicas allocation at nodes of WAN.

Let $T(X_r, Y_r)$ be the minimal average delay per packet in WAN in which values of channel capacities are given by X_r and traffic requirements are given by Y_r (depending on server replica allocation). $T(X_r, Y_r)$ can be obtained by solving a multi-commodity flow problem in the network [8]. Let $U(Y_r)$ be the server cost and let $d(X_r)$ be the capacity cost. Let $Q(X_r, Y_r)$ be linear combination of the total average delay per packet, the server cost and the capacity cost

$$Q(X_r, Y_r) = \alpha T(X_r, Y_r) + \beta U(Y_r) + \rho d(X_r) \tag{4}$$

where α, β and ρ are the positive coefficients; $\alpha, \beta, \rho \in [0, 1], \alpha + \beta + \rho = 1$.

Let B be the budget (maximal feasible leasing capacity cost of channels) of WAN. Then, the considered server replication and topology assignment problem in WAN can be formulated as follows.

$$\min_{(X_r, Y_r)} Q(X_r, Y_r) \tag{5}$$

subject to

$$(X_r, Y_r) \in \mathfrak{R} \tag{6}$$

$$d(X_r) = \sum_{x_j \in X_r} x_j^i d_j^i \leq B \tag{7}$$

3 Calculation Scheme of the Branch and Bound Algorithm

Assuming that $LK_k = 1$ for $k=1, \dots, K$ and $C^i = \bar{C}^i$ for $i=1, \dots, b$, the problem (5-7) is resolved itself into the “host allocation, capacity and flow assignment problem”. Since the host allocation, capacity and flow assignment problem is NP-complete [6], [8] then the problem (5-7) is also NP-complete as more general. Then, the branch and bound method can be used to construct the exact algorithm. Starting with the selection $(X_1, Y_1) \in \mathfrak{R}$ we generate a sequence of selections (X_s, Y_s) . Each selection (X_s, Y_s) is obtained from a certain selections (X_r, Y_r) of the sequence by complementing one variable x_j^i (or y_{kh}) by another variable from W^i (or $\{y_{km} : m \in M_k \text{ and } m \neq h\}$).

So, for each selection (X_r, Y_r) we constantly fix a subset $F_r \in (X_r, Y_r)$ and momentarily fix a set F_r^t . The variables in F_r are constantly fixed and represent the path from the initial selection (X_1, Y_1) to the selection (X_r, Y_r) . Each momentarily fixed variable in F_r^t is the variable abandoned during the backtracking process. Variables, which do not belong to F_r or F_r^t are called free in (X_r, Y_r) . There are two important elements in branch and bound method: testing operation (lower bound of the criterion function) and branching rules. Then, in the next section of the paper, the testing operation and choice operation are proposed.

The lower bound LB_r and branching rules are calculated for each selection (X_r, Y_r) . The lower bound is calculated to check if the “better” solution (selection (X_s, Y_s)) may be found. If the testing is negative, we abandon the considered selection (X_r, Y_r) and backtrack to the selection (X_p, Y_p) from which selection (X_r, Y_r) was generated. The basic task of the branching rules is to find the variables for complementing to generate a new selection with the least possible value of the criterion function. The detailed description of the calculation scheme of branch and bound method may be found in the paper [9].

4 Lower Bound

To find the lower bound LB_r of the criterion function (4) we reformulate the problem (5-7) in the following way. We assume that the variables x_j^i and y_{kh} are continuous

variables, we omit the constraint (2) and approximate the discrete cost-capacity curves (given by the set C^i) with the lower linear envelope. In this case, the constraint (7) may be relaxed by the constraint $\sum d^i c^i \leq B$, where $d^i = \min_{x_j^i \in W^i} (d_j^i / c_j^i)$

and c^i is the capacity of the channel i (continuous variable). To easy find the lower bound we create the model of the WAN in the following way. We add to the considered network $2K$ new artificial nodes, numbered from $n+1$ to $n+2K$. The artificial nodes $n+k$ and $n+K+k$ correspond to the k -th host. Moreover we add to the network directed artificial channels $\langle n+k, m \rangle$, $\langle m, n+K+k \rangle$, $\langle n+K+k, n+k \rangle$, such that $m \in M_k$ and $y_{kh} \in ZY' \cup ZY''$. The capacities of the new artificial channels are following: $c(n+k, m) = \infty$, $c(m, n+K+k) = \infty$, $c(n+K+k, n+k) = \sum_{h=1}^n u_{kh}$.

The leasing costs of all artificial channels are equal to zero.

Then, the lower bound LB_r of minimal value of the criterion function $Q(X_s, Y_s)$ for every possible successor (X_s, Y_s) generated from (X_r, Y_r) may be obtained by solving the following optimization problem:

$$LB_r = \min_{\underline{f}} \left(2\sqrt{\frac{\alpha}{\gamma}} \sum_{i: x_j^i \in X_r - F_r} \sqrt{d^i f_i} + \frac{\alpha}{\gamma} \sum_{x_j^i \in F_r} \frac{f_i}{x_j^i c_j^i - f_i} + \rho \sum_{i: x_j^i \in X_r - F_r} d^i f_i + \right. \\ \left. + \rho \sum_{x_j^i \in F_r} x_j^i d_j^i + \beta \sum_{y_{kn} \in Y_r - F_r} \frac{f(n, N+K+k) + f(N+k, n)}{\sum_{m=1}^N (\tilde{r}_{km}' + \check{r}_{km}')} u_{kn} + \beta \sum_{y_{kn} \in F_r} u_{kn} y_{kn} \right) \quad (8)$$

subject to

$$f_i \leq x_j^i c^i \quad \text{for } x_j^i \in X_r - F_r \quad (9)$$

$$f_i \leq x_j^i c_j^i \quad \text{for } x_j^i \in F_r \quad (10)$$

$$f_i \leq c_{\max}^{ir} \quad \text{for each } x_j^i \in X_r - F_r \quad (11)$$

$$\sum_{e=1}^{N+K} f^{nm}(a, e) - \sum_{e=1}^{N+K} f^{nm}(e, a) = \begin{cases} \tilde{r}_{nm} & \text{if } a = n \text{ and } n, m \leq N \\ -\tilde{r}_{nm} & \text{if } a = m \text{ and } n, m \leq N \\ \hat{r}_{pm}' & \text{if } a = m \text{ and } e = N + p \\ \check{r}_{pm}' & \text{if } a = N + p \text{ and } e = m \end{cases} \quad (12)$$

where \tilde{r}_{nm} is the average flow rate directed from node n to node m , \tilde{r}_{pm} is the average flow rate directed from node m to p -th server, \tilde{r}'_{pm} is the average flow rate directed from p -th server to node m and $c_{\max}^{ir} = \max_{x_j^i \in W^i - F_r^i} c_j^i$.

The solution of problem (8–12) gives the lower bound LB_r . To solve the above reformulated problem we can use an efficient Flow Deviation method [6], [8].

5 Branching Rules

The purposes of branching rules is to find the normal variable from the selection (X_r, Y_r) for complementing and generating a successor (X_s, Y_s) of the selection (X_r, Y_r) with the least possible value of the criterion function (4). We can choose a variable x_j^i or a variable y_{kh} .

Complementing variables x_j^i causes the capacity change in channel i . Then, the values of $T(X_r, Y_r)$ and $d(X_r)$ changes, and the value of $U(Y_r)$ does not change. Then, in this case we can use the criterion given in the form of the following theorem.

Theorem 1. Let $(X_r, Y_r) \in \mathfrak{R}$. If the selection (X_s, Y_s) is obtained from the selection (X_r, Y_r) by complementing the variable $x_j^i \in X_r$ by the variable $x_l^i \in X_s$ then $Q(X_s, Y_s) \leq \Delta_{jl}^i$, where

$$\Delta_{jl}^i = \begin{cases} Q(X_r, Y_r) - \frac{\alpha}{\gamma} \left(\frac{f_i}{c_j^i - f_i} - \frac{f_i}{c_l^i - f_i} \right) - \rho(x_j^i d_j^i - x_l^i d_l^i) & \text{if } x_j^i c_j^i > f_i \\ Q(X_r, Y_r) - \frac{\alpha}{\gamma} \left(\frac{f_i}{c_j^i - f_i} - \frac{f_i}{c_{\max}^{ir} - f_i} \right) - \rho(x_j^i d_j^i - x_l^i d_l^i) & \text{otherwise} \end{cases} \tag{13}$$

f_r^i is the flow in the i -th channel obtained by solving the multicommodity flow problem for network topology and channel capacities given by the selection X_r and γ is the total average packet rate from external sources.

To formulate the choice criterion on variables y_{kh} we use the following theorem.

Theorem 2. Let $(X_r, Y_r) \in \mathfrak{R}$. If the selection (X_s, Y_s) is obtained from the selection (X_r, Y_r) by complementing the variable $y_{kh} \in Y_r$ by the variable $y_{km} \in Y_s$ then $Q(X_s, Y_s) \leq \delta_{hm}^k$, where

$$\delta_{hm}^k = \begin{cases} \frac{\alpha}{\gamma} \sum_{x_j^i \in X_r} \frac{\tilde{f}^i}{x_j^i c_j^i - \tilde{f}^i} + \beta(U(Y_r) - u_{kh} + u_{km}) + \rho d(X_r) & \text{if } \tilde{f}^i < x_j^i c_j^i \text{ for } x_j^i \in X_r \\ \infty & \text{otherwise} \end{cases} \quad (14)$$

$\tilde{f}^i = f_r^i - f_{ik}^i + f_{ik}''$, f_{ik}^i and f_{ik}'' are parts of flow in i -th channel; f_{ik}^i corresponds to the packets sent from all users to those replica of k -th server which is allocated at node h (before complementing) and from this replica to all users. f_{ik}'' corresponds to the packets exchanged between all users and replica after reallocating replica from node h to node m ; u_{kh} is the cost of connecting the server k at node h .

Let $E_r = (X_r \cup Y_r) - F_r$, and let G_r be the set of all reverse variables of normal variables, which belong to the set E_r . We want to choose a normal variable the complementing of which generates a successor with the possible least value of criterion (4). We should choose such pairs $\{(y_{kh}, y_{km}): y_{kh} \in E_r, y_{km} \in G_r\}$ or $\{(x_j^i, x_l^i): x_j^i \in E_r, x_l^i \in G_r\}$ for which the value of criterion δ_{hm}^k or Δ_{jl}^i is minimal.

6 Computational Results

The presented exact algorithm was implemented in C++ code. Extensive numerical experiments have been performed with this algorithm for many different network topologies and for many sets of possible server’s replica locations and connecting costs. The experiments were conducted with two main purposes in mind: first, to test the computational efficiency of an algorithm and second, to examine the impact of various parameters on solutions.

Let $NB = ((B - D_{\min}) / (D_{\max} - D_{\min})) \cdot 100\%$ be the normalized budget in percentage. D_{\max} is the maximal capacity cost of the network and D_{\min} is the minimal

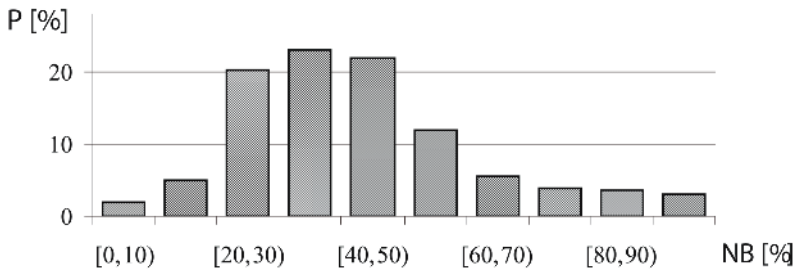


Fig. 1. The dependence of P on normalized budget NB

capacity cost - problem (5-7) has no solution for $B < D_{\min}$. Normalized budget let us compare the results obtained for different wide area network topologies and for different replica locations. Let $P(u, v)$, in percentage, be the arithmetic mean of the relative number of iterations for $NB \in [u, v]$ calculated for all considered network topologies and for different replica locations. Fig. 1 shows the dependency of P on divisions $[0\%, 10\%), [10\%, 20\%), \dots, [90\%, 100\%]$ of normalized budget NB . It follows from Fig. 1 that the exact algorithm is especially effective from computational point of view for $NB \in [0\%, 20\%] \cup [70\%, 100\%]$.

The typical dependence of the optimal value of Q on budget B for different values of coefficients α, β, ρ is presented in the Fig. 2. It follows from Fig. 2 that there exists such budget B^* , that the problem (5-7) has the same solution for each B greater than or equal to B^* .

Conclusion 1. In the problem (5-7), for $B \geq B^*$ the constraint (7) may be substituted by constraint $d(X_r) \leq B^*$.

This observation shows that the influence of the building cost (budget) on the optimal solution of the problem (5-7) is very limited for greater values of budget B .

Let optimal value of Q obtained by solving the problem (5-7) is following:

$$Q = \alpha T^{opt} + \beta U^{opt} + \rho d^{opt}.$$

Then, very interesting problem is to study the dependence of $T^{opt}, U^{opt}, d^{opt}$ on coefficients α, β, ρ .

The dependence of the average delay per packet T^{opt} on coefficient α has been examined. In the Fig. 3 the typical dependence of T^{opt} [microsecond] on α is presented for different values of budget B . The typical dependence of the optimal value of criterion U^{opt} on coefficient β for different values of the network budget B is presented in the Fig. 4, and the typical dependence of the optimal value d^{opt} on the coefficient ρ for different values of B is presented in the Fig. 5.

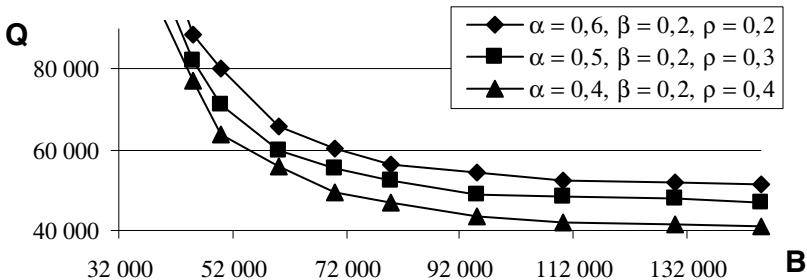


Fig. 2. Dependence of Q on the budget B for different values of coefficients α, β, ρ

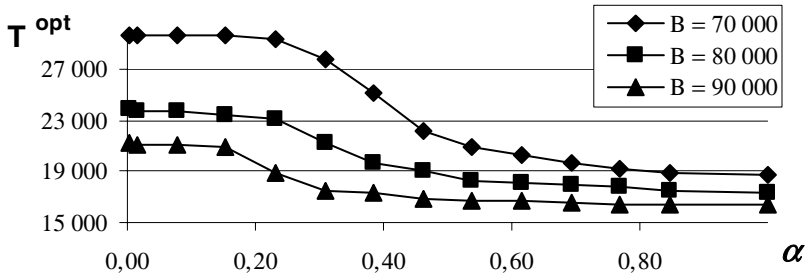


Fig. 3. The typical dependence of T^{opt} on coefficient α for different values of B

It follows from the numerical experiments and from the Fig. 3, 4, 5 that the functions $T^{opt}(\alpha)$, $U^{opt}(\beta)$ and $d^{opt}(\rho)$ are decreasing.

Conclusion 2. Values of coefficients α, β, ρ have the impact on the values of T^{opt}, U^{opt} and d^{opt} in the optimal value of criterion Q , obtained by solving the problem (5-7).

The conclusions formulated above are important from practical point of view. They allow to specify the designing assumptions on coefficients α, β and ρ . During the design process, one of the criteria may be more important and the other little less important for designer. When we want one of the three criteria to be the least we have to determine the biggest coefficient for this criterion. Network designers can indicate which part of the criterion Q (i.e. T^{opt} or U^{opt} or d^{opt}) is the most important. It follows from Conclusion 2 that it may be done by appropriate determine the value of coefficient α, β and ρ .

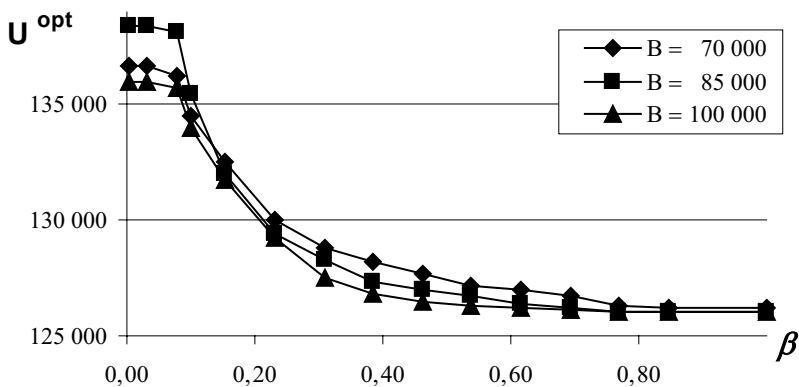


Fig. 4. The typical dependence of U^{opt} on coefficient β for different values of B

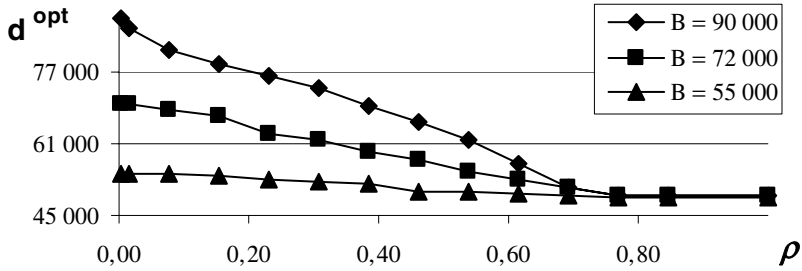


Fig. 5. The typical dependence of d^{opt} on coefficient ρ for different values of B

7 Conclusion

In the paper an exact algorithm for solving the three-criteria servers replication and network topology assignment problem in WAN is presented. The considered problem is far more general than the similar problems presented in the literature. Considering two different kinds of cost (capacity cost and server costs) is very important from practical point of view. Some properties of the considered problem have been discovered and formulated as conclusions. In our opinion, in practice the most important is conclusion 2, because it helps to make designing assumptions.

This work was supported by a research project of The Polish State Committee for Scientific Research in 2005-2007.

References

1. Pioro, M., Medhi, D.: Routing, Flow, and Capacity Design in Communication and Computer Networks, Elsevier, Morgan Kaufmann Publishers, San Francisco (2004)
2. Walkowiak, K.: QoS Dynamic Routing in Content Delivery Network, Lectures Notes in Computer Science, Vol. 3462 (2005), 1120-1132
3. Markowski, M., Kasprzak, A.: The web replica allocation and topology assignment problem in wide area networks: algorithms and computational results, Lecture Notes in Computer Science, Vol. 3483 (2005), 772-781
4. Radoslavov, P., Govindan, R., Estrin, D.: Topology-Informed Internet Replica Placement, Computer Communications, Volume: 25, Issue: 4 (2002), 384-392
5. Qiu, L., Padmanabhan, V. N., Voelker, G. M.: On the Placement of Web Server Replicas, Proc. of 20th IEEE INFOCOM, Anchorage, USA (2001), 1587-1596
6. Kasprzak, A.: Topological Design of the Wide Area Networks. Wroclaw University of Technology Press, Wroclaw (2001)
7. Kang, C. G., Tan, H. H.: Combined channel allocation and routing algorithms in packed switched networks, Computer Communications, vol. 20 (1997), 1175-1190
8. Fratta, L., Gerla, M., Kleinrock, L.: The Flow Deviation Method: an Approach to Store-and-Forward Communication Network Design. Networks 3 (1973), 97-133
9. Wolsey, L.A.: Integer Programming. Wiley-Interscience, New York (1998)

An Efficient Multicast Tree with Delay and Delay Variation Constraints*

Moonseong Kim¹, Young-Cheol Bang², Jong S. Yang³, and Hyunseung Choo¹

¹ School of Information and Communication Engineering,
Sungkyunkwan University, 440-746, Suwon, Korea
Tel.: +82-31-290-7145

{moonseong, choo}@ece.skku.ac.kr

² Department of Computer Engineering,
Korea Polytechnic University, 429-793, Gyeonggi-Do, Korea
Tel.: +82-31-496-8292

ybang@kpu.ac.kr

³ Korea Institute of Industrial Technology Evaluation and Planning, Seoul, Korea
Tel.: +82-42-860-6333
yjs@mail.itep.re.kr

Abstract. With the rapid evolution of real time multimedia applications like audio/video conferencing, interactive distributed games and real time remote control system, a certain Quality of Service (QoS) needs to be guaranteed in underlying networks. Multicast routing algorithms should support the required QoS. There are two important QoS parameters, bounded delay and delay variation, that need to be guaranteed in order to support the real time multimedia applications. Here we solve Delay and delay Variation Bounded Multicast Tree (DVBMT) problem which has been proved to NP-complete. In this paper, we propose an efficient algorithm for DVBMT. The performance enhancement is up to about 21.7% in terms of delay variation as compared to the well-known algorithm, KBC [9].

1 Introduction

New communication services involving multicast communications and real time multimedia applications are becoming prevalent. The general problem of multicasting is well studied in the area of computer networks and algorithmic network theory. In multicast communications, messages are sent to multiple destinations that belong to the same multicast group. These group applications demand a certain amount of reserved resources to satisfy their Quality of Service (QoS) requirements such as end-to-end delay, delay variation, cost, loss, throughput, etc.

* This research was supported by the Ministry of Information and Communication, Korea under the Information Technology Research Center support program supervised by the Institute of Information Technology Assessment, IITA-2005-(C1090-0501-0019) and the Ministry of Commerce, Industry and Energy under Next-Generation Growth Engine Industry. Dr. Choo is the corresponding author and Dr. Bang is the co-corresponding author.

The multicast tree problem can be modelled as the Steiner tree problem which is NP-complete. A lot of heuristics [1] [2] [3] that construct low-cost multicast routing based on the problem are proposed. Not only the tree cost as a measure of bandwidth efficiency is one of the important factors, but also networks supporting real-time traffic are required to receive messages from source node in a limited amount of time. If the messages cannot reach to the requested destination nodes within the time, they may be useless information. Therefore, it can be an important factor to guarantee an upper bound on the end-to-end delay from the source to each destination. This is called the multicast end-to-end delay problem [4].

There is another important factor to consider real-time application in the multicast network. During video-conferencing, the person who is talking has to be heard by all participants at the same time. Otherwise, the communication may lose the feeling of an interactive screen-to-screen discussion. Moreover, the situation of the on-line video gaming may have the problem that the characters who are in the game must be moved simultaneously. All these problems can be explained that the delay variation has to be kept within a restricted time. They are all related to the multicast delay variation problem [5] [6].

In this paper, we study the delay variation problem under the upper bound on the multicast end-to-end delay. We propose an efficient algorithm in comparison with current algorithms known as the best algorithms so far. The proposed algorithm has the better performance than current algorithms have in terms of the multicast delay variation. The rest of the paper is organized as follows. In Section 2, we state the network model for the multicast routing, the problem formulations, and the previous algorithms. Section 3 presents the details of the proposed algorithms. Then, we evaluate the proposed algorithms by the computer simulation, in Section 4. Finally, Section 5 concludes this paper.

2 Preliminaries

2.1 Network Model

We consider that a computer network is represented by a directed graph $G = (V, E)$ with n nodes and l links, where V is a set of nodes and E is a set of links, respectively. Each link $e = (i, j) \in E$ is associated with two parameters, namely link cost $c(e) \geq 0$ and link delay $d(e) \geq 0$. The delay of a link, $d(e)$, is the sum of the perceived queueing delay, transmission delay, and propagation delay. We define a path as sequence of links such that $(u, i), (i, j), \dots, (k, v)$, belongs to E . Let $P(u, v) = \{(u, i), (i, j), \dots, (k, v)\}$ denote the path from node u to node v . For given a source node $s \in V$ and a destination node $d \in V$, $(2^{s \rightarrow d}, \infty)$ is the set of all possible paths from s to d .

$$(2^{s \rightarrow d}, \infty) = \{P_k(s, d) \mid \text{all possible paths from } s \text{ to } d, \forall s, d \in V, \forall k \in \Lambda\} \quad (1)$$

where Λ is an index set. The path cost of P is given by $\phi_C(P) = \sum_{e \in P} c(e)$ and the path delay of P is given by $\phi_D(P) = \sum_{e \in P} d(e)$. $(2^{s \rightarrow d}, \Delta)$ is the set

of paths from s to d for which the end-to-end delay is bounded by Δ . Therefore $(2^{s \rightarrow d}, \Delta) \subseteq (2^{s \rightarrow d}, \infty)$.

For the multicast communications, messages need to be delivered to all receivers in the set $M \subseteq V \setminus \{s\}$ which is called the multicast group, where $|M| = m$. The path traversed by messages from the source s to a multicast receiver, m_i , is given by $P(s, m_i)$. Thus multicast routing tree can be defined as $T(s, M)$, the Steiner tree of $\bigcup_{m_i \in M} P(s, m_i)$, and the messages are sent from s to M through $T(s, M)$. The tree delay is $\phi_D(T) = \max\{\phi_D(P(s, m_i)) \mid \forall P(s, m_i) \subseteq T, \forall m_i \in M\}$.

The multicast delay variation, $\phi_\delta(T)$, is the maximum difference between the end-to-end delays along the paths from the source to any two destination nodes.

$$\phi_\delta(T) = \max\{|\phi_D(P(s, m_i)) - \phi_D(P(s, m_j))|, \forall P \subseteq T, \forall m_i, m_j \in M, i \neq j\} \quad (2)$$

The issue defined and discussed in [5], initially, is to minimize multicast delay variation under multicast end-to-end delay constraint. The authors referred to this problem as Delay- and delay Variation-Bounded Multicast Tree (DVBMT) problem. The DVBMT problem is to find the tree that satisfies

$$\min\{\phi_\delta(T_\alpha) \mid \forall P(s, m_i) \in (2^{s \rightarrow m_i}, \Delta), \forall P(s, m_i) \subseteq T_\alpha, \forall m_i \in M, \forall \alpha \in \Lambda\} \quad (3)$$

where T_α denotes any multicast tree spanning $M \cup \{s\}$, and is known to be NP-complete [5].

2.2 Previous Algorithms

The DVBMT problem has been introduced in [5]. Rouskas and Baldine proposed Delay Variation Multicast Algorithm (DVMA), finds a Multicast Tree spanning the set of multicast nodes. DVMA works on the principal of finding the k^{th} shortest paths to the concerned nodes. If these paths do not satisfy a delay variation bound δ , longer paths are found. The complexity of DVMA is $O(klmn^4)$, where k and l are the largest value among the numbers of the all appropriate paths between any two nodes under Δ , $|M| = m$, and $|V| = n$. DVMA is a high time complexity does not fit in modern high speed computer network environment.

Sheu and Chen proposed Delay and Delay Variation Constraint Algorithm (DDVCA) [7] based on Core Based Trees (CBT) [8]. Since DDVCA is meant to search as much as possible for a multicast tree with a smaller multicast delay variation under the multicast end-to-end delay constraint Δ , DDVCA in picking a central node has to inspect whether that central node will violate the multicast end-to-end delay constraint Δ . In spite of DVMA's smart performance in terms of the multicast delay variation, its time complexity is very high. But DDVCA has a much lower time complexity $O(mn^2)$ and has a satisfactory performance.

Kim, *et al.* have recently proposed a heuristic algorithm based on CBT like DDVCA [9], their algorithm hereafter referred to as KBC. DDVCA overlooked a portion of the core selection. The selection of a core node over several candidates (possible core nodes) is randomly selected among candidate nodes. Meanwhile

KBC investigates candidate nodes to select the better node with the same time complexity of DDVCA. KBC obtains the better minimum multicast delay variation than DDVCA.

2.3 Weighted Factor Algorithm (WFA)

Kim, *et al.* have recently proposed a unicast routing algorithm, their algorithm hereafter referred to as WFA [10], which is probabilistic combination of the link cost and delay and its time complexity is $O(Wl + Wn \log n)$, where $\{\omega_\alpha\}$ is set of weights and $|\{\omega_\alpha\}| = W$. The authors investigated the efficiency routing problem in point-to-point connection-oriented networks with a QoS. They formulated the new weight parameter that simultaneously took into account both the link cost and delay. The Least Delay path (P_{LD}) cost is relatively more expensive than the Least Cost path (P_{LC}) cost, and moreover, $\phi_D(P_{LC})$ is relatively higher than $\phi_D(P_{LD})$. The unicast routing algorithm, WFA, is quite likely a performance of k^{th} shortest path algorithm. The weight ω plays an important role in combining the two independent measures. If the ω is nearly to 0, then the path delay is low as in Fig. 1. Otherwise the path cost is low. Thus, the efficient routing path can be determined once ω is selected. Our proposed algorithm uses WFA for construction of multicast tree. Since a multicast tree is union of paths for every multicast member, we can select suitable weight for every multicast member.

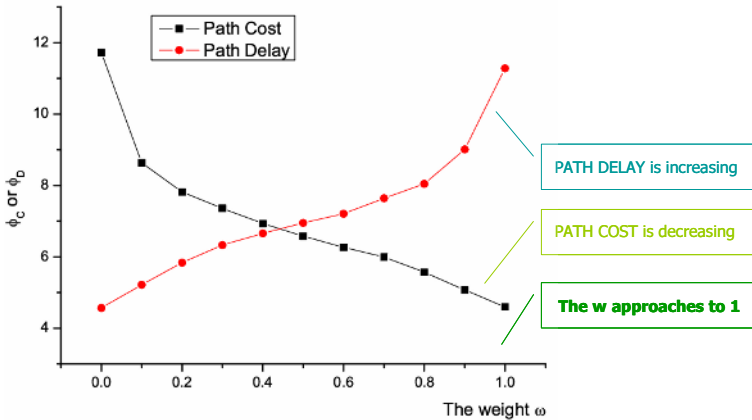


Fig. 1. Relationship between path delay and path cost from [10] ($P_c: 0.3, |V|: 100$)

3 The Proposed Algorithm

We now present algorithm to construct a multicast tree satisfying constraints (3) for the given value of the path delay Δ . We assume that complete information

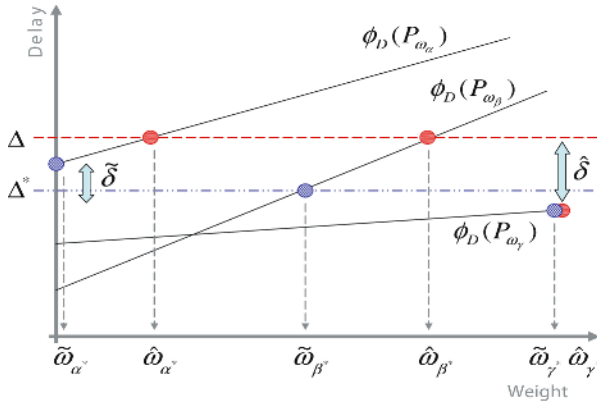


Fig. 2. Weight selection in proposed algorithm

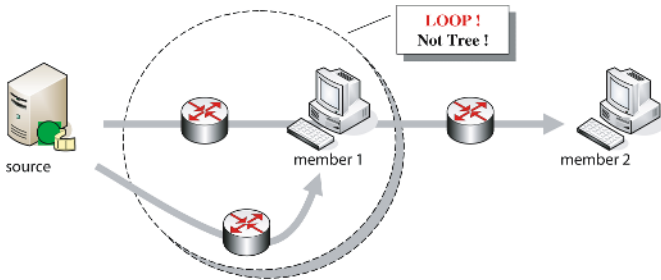


Fig. 3. Scenario illustrating loop creation in $G = P(s, m_1) \cup P(s, m_2)$

regarding the network topology is stored locally at source node s , making it possible to determine the multicast tree at the source node itself. This information may be collected and updated using an existing topology broadcast algorithm.

Our objective is to obtain the feasible tree of minimum delay variation for the given value of Δ . We introduce a heuristic algorithm for DVMT. WFA can check various paths with path delay for each ω as Fig. 1. For each multicast member, the proposed algorithm uses WFA and finds a pertinence weight with Δ . Let $G = (V, E)$ be a given network and $M = \{ m_\alpha, m_\beta, m_\gamma \} \subset V$ be a multicast group. As indicated in Fig. 2, we select $\hat{\omega}_i$ such that $\max\{ \phi_D(P_{\hat{\omega}_i}) \mid \forall P_{\omega_i} \in (2^{s \rightarrow m_i}, \Delta) \forall \omega_i \in W \}$. If the difference between $\phi_D(P_{\hat{\omega}_\alpha})$ and $\phi_D(P_{\hat{\omega}_\gamma})$ is big, then a potential delay variation $\hat{\delta}$ is large. So we have to reduce $\hat{\delta}$. Let Δ^* be a virtual delay boundary in place of Δ . Hence we have $\Delta^* = \sum_{m_i \in M} \phi_D(P_{\hat{\omega}_i}) / |M|$ as the arithmetic mean for $\phi_D(P_{\hat{\omega}_i})$. And we select $\tilde{\omega}_i$ such that $\min\{ | \phi_D(P_{\tilde{\omega}_i}) - \Delta^* |, \forall P_{\omega_i} \in (2^{s \rightarrow m_i}, \Delta) \}$. Intuitively, a potential delay variation $\tilde{\delta}$ may be smaller than $\hat{\delta}$ (i.e., $\tilde{\delta} \leq \hat{\delta}$).

Let $G^* = \bigcup_{m_i \in M} P_{\tilde{\omega}_i}(s, m_i)$ be a connected subgraph of G . See Fig. 3, G^* must be not tree. We have to find the minimal spanning tree, T^* , of G^* . If

there are several minimal spanning trees, pick an arbitrary one. And then, the proposed algorithm constructs a multicast tree, $T(s, M)$, from T^* by deleting links in T^* , if necessary, so that all the leaves in $T(s, M)$ are multicast members.

4 Performance Evaluation

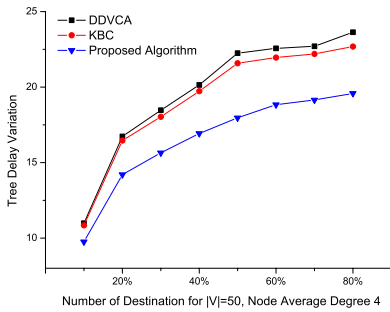
4.1 Random Real Network Topology for the Simulation

Random graphs of the acknowledged model represent different kinds of networks, communication networks in particular. There are many algorithms and programs, but the speed is usually the main goal, not the statistical properties. In the last decade the problem was discussed, for examples, by B. M. Waxman (1993) [11], M. Doar (1993, 1996) [12][13], C.-K. Toh (1993) [14], E. W. Zegura, K. L. Calvert, and S. Bhattacharjee (1996) [15], K. L. Calvert, M. Doar, and M. Doar (1997) [16], R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal (2000) [17]. They have presented fast algorithms that allow the generation of random graphs with different properties, in particular, these are similar to real communication networks. However, none of them have discussed the stochastic properties of generated random graphs. A. S. Rodionov and H. Choo [18] have formulated two major demands for the generators of random graph: attainability of all graphs with required properties and uniformity of distribution. If the second demand is sometimes difficult to prove theoretically, it is possible to check the distribution statistically. The method uses parameter P_e , the probability of link existence between any node pair. We use the method by Rodionov and Choo.

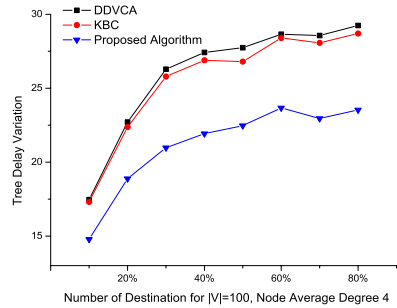
4.2 Simulation Results

We now describe some numerical results with which we compare the performance of the proposed schemes. We generate 100 different networks for each size of 50, 100, and 200. Each node in network has the node average degree 4 or $P_e = 0.3$. The proposed algorithm is implemented in C . We randomly select a source node. The destination nodes are picked uniformly from the set of nodes in the network topology (excluding the nodes already selected for the destination). Moreover, the destination nodes in the multicast group, M , are occupied 10% ~ 80% of the overall nodes on the network. The delay bound Δ value in our computer experiment is set to be 1.5 times the minimum delay between the source node and the farthest destination node [7]. We simulate 1000 times ($10 \times 100 = 1000$) for each $|V|$.

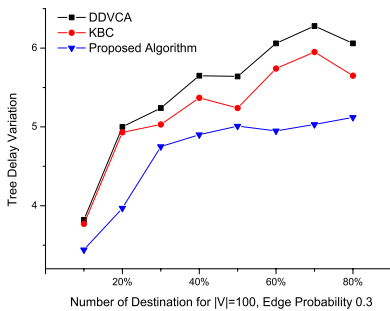
For the performance comparison, we implement DDVCA and KBC in the same simulation environments. As indicated in Fig. 4, it is easily noticed that the proposed algorithm is always better than others. KBC is a little better than DDVCA. Since our algorithm specially regards the weight selection, its delay



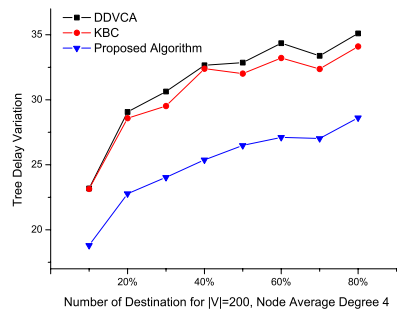
(a) $|V| = 50$ with average degree 4



(b) $|V| = 100$ with average degree 4



(c) $|V| = 100$ with $P_e = 0.3$



(d) $|V| = 200$ with average degree 4

Fig. 4. Tree delay variations

variation cannot help being good. The enhancement is up to about 16.1% ~ 21.7% ($|V| = 200$) in terms of delay variation for KBC.

5 Conclusion

We studied the problem of constructing minimum delay variation multicast tree that satisfies the end-to-end delay bound and delay variation bound, which is called as DVBMT, and has been proved to be NP-complete. We proposed algorithm using expected multiple paths. The expected multiple paths are obtained by WFA which is introduced in [10]. WFA is efficiently combining two independent measures, the link cost and delay. The weight ω plays an important role in combining the two measures. Our algorithm finds suitable weight ω for each destination member, and they construct the multicast tree for DVBMT. The efficiency of our algorithm is verified through the performance evaluations and the enhancements are 16.1% ~ 21.7% in terms of the multicast delay variation. Also, the time complexity is $O(Wml + Wmn \log n)$ which is comparable to one of previous works.

References

1. Y.-C. Bang and H. Choo, "On multicasting with minimum costs for the Internet topology," Springer-Verlag Lecture Notes in Computer Science, vol. 2400, pp. 736-744, August 2002.
2. L. Kou, G. Markowsky, and L. Berman, "A fast algorithm for Steiner trees," *Acta Informatica*, vol. 15, pp. 141-145, 1981.
3. H. Takahashi and A. Matsuyama, "An Approximate Solution for the Steiner Problem in Graphs," *Mathematica Japonica*, vol. 24, no. 6, pp. 573-577, 1980.
4. V. P. Kompella, J. C. Pasquale, and G. C. Polyzos, "Multicast routing for multimedia communication," *IEEE/ACM Trans. Networking*, vol. 1, no. 3, pp. 286-292, June 1993.
5. G. N. Rouskas and I. Baldine, "Multicast routing with end-to-end delay and delay variation constraints," *IEEE JSAC*, vol. 15, no. 3, pp. 346-356, April 1997.
6. M. Kim, Y.-C. Bang, and H. Choo, "On Estimation for Reducing Multicast Delay Variation," Springer-Verlag Lecture Notes in Computer Science, HPCCC 2005, vol. 3726, pp. 117-122, September 2005.
7. P.-R. Sheu and S.-T. Chen, "A Fast and Efficient Heuristic Algorithm for the Delay- and Delay Variation-Bounded Multicast Tree Problem," *Computer Communications*, vol. 25, no. 8, pp. 825-833, 2002.
8. A. Ballardie, B. Cain, Z. Zhang, "Core Based Trees (CBT version 3) Multicast Routing protocol specification," Internet Draft, IETF, August 1998.
9. M. Kim, Y.-C. Bang, and H. Choo, "Efficient Algorithm for Reducing Delay Variation on Bounded Multicast Trees," Springer-Verlag Lecture Notes in Computer Science, Networking 2004, vol. 3090, pp. 440-450, September 2004.
10. M. Kim, Y.-C. Bang, and H. Choo, "On Algorithm for Efficiently Combining Two Independent Measures in Routing Paths," Springer-Verlag Lecture Notes in Computer Science, ICCSA 2005, vol. 3483, pp. 989-998, May 2005.
11. B. W. Waxman, "Routing of multipoint connections," *IEEE JSAC*, vol. 6, no. 9, pp. 1617-1622, December 1988.
12. M. Doar, "Multicast in the ATM environment," Ph.D dissertation, Cambridge University, Computer Lab., September 1993.
13. M. Doar, "A Better Mode for Generating Test Networks," *IEEE Proc. GLOBE-COM'96*, pp. 86-93, 1996.
14. C.-K. Toh, "Performance Evaluation of Crossover Switch Discovery Algorithms for Wireless ATM LANs," *IEEE Proc. INFOCOM96*, pp. 1380-1387, 1996.
15. E. W. Zegura, K. L. Calvert, and S. Bhattacharjee, "How to model an Internet network," *IEEE Proc. INFOCOM96*, pp. 594-602, 1996.
16. K. L. Calvert, M. Doar, and M. Doar, "Modelling Internet Topology," *IEEE Communications Magazine*, pp. 160-163, June 1997.
17. R. Kumar, P. Raghavan, S. Rajagopalan, D Sivakumar, A. Tomkins, and E. Upfal, "Stochastic models for the Web graph," *Proc. 41st Annual Symposium on Foundations of Computer Science*, pp. 57-65, 2000.
18. A. S. Rodionov and H. Choo, "On Generating Random Network Structures: Connected Graphs," Springer-Verlag Lecture Notes in Computer Science, vol. 3090, pp. 483-491, August 2004.

Algorithms on Extended (δ, γ) -Matching*

Inbok Lee¹, Raphaël Clifford², and Sung-Ryul Kim^{3,**}

¹ King's College London, Department of Computer Science,
London WC2R 2LS, UK

`inboklee@gmail.com`

² University of Bristol, Department of Computer Science, UK
`raphael@clifford.net`

³ Konkuk University, Division of Internet & Media and CAESIT,
Seoul, Republic of Korea
`kimsr@konkuk.ac.kr`

Abstract. Approximate pattern matching plays an important role in various applications, such as bioinformatics, computer-aided music analysis and computer vision. We focus on (δ, γ) -matching. Given a text T of length n , a pattern P of length m , and two parameters δ and γ , the aim is to find all the substring $T[i, i + m - 1]$ such that (a) $\forall 1 \leq j \leq m$, $|T[i+j-1]-P[j]| \leq \delta$ (δ -matching), and (b) $\sum_{1 \leq j \leq m} |T[i+j-1]-P[j]| \leq \gamma$ (γ -matching). In this paper we consider three variations of (δ, γ) -matching: *amplified matching*, *transposition-invariant matching*, and *amplified transposition-invariant matching*. For each problem we propose a simple and efficient algorithm which is easy to implement.

1 Introduction

Approximate pattern matching plays an important role in various applications, such as bioinformatics, computer-aided music analysis and computer vision. In real world, patterns rarely appear exactly in the text due to various reason. Hence we are interested in finding *approximate* occurrence of the pattern from the text. When we handle numeric data, approximate occurrences mean small differences between the text and the pattern.

We focus on (δ, γ) -matching. Informally (δ, γ) -matching refers to the problem of finding all the substrings of the text where each character differs at most δ and the sum of the differences is equal to or smaller than γ . Even though (δ, γ) -matching is one way of modelling approximate matching, we need to handle another approximation in addition to that. We consider three variations of (δ, γ) -matching. Informally, our problems are as follows:

- [AMPLIFIED MATCHING] refers to finding all the substrings in T where each character of P is multiplied by an arbitrary number.

* This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

** Contact author.

- [TRANSPPOSITION-INVARIANT MATCHING] refers to finding all the substrings in T where each character of P is added by an arbitrary number.
- [AMPLIFIED TRANSPPOSITION-INVARIANT MATCHING] refers to the combination of two. We want to find all the substrings in T where each character of P is first multiplied as in amplified matching, then added by an arbitrary number.

Another way to explain these problems is to consider a linear transformation $y = \alpha x + \beta$. If $\alpha = 0$, it is transposition-invariant matching. If $\beta = 0$, then it is amplified matching. Otherwise, it is amplified transposition-invariant matching.

These problems happen in real world applications. In music analysis, a simple motive can be transformed into different variations, changing frequencies or duration of notes. If we represent musical data in numerical form, the original pattern is amplified or added by a constant number. J.S. Bach’s Goldberg Variations and L. Beethoven’s Diabelli Variations are famous examples of extending a simple motive into a great masterpiece. Hence finding the occurrence of the simple motive in varied forms plays an important role in music analysis. Figure 1 shows an example. The pattern is presented in (a). In (b), each character in P is multiplied by two. They are represented by dashed circles. Now assuming $\delta = 1$ and $\gamma = 2$, there is an occurrence of amplified matching in the text (black dots). In (c), each character in P is added by one. The dashed circles and black dots are the same as in (b) and it shows an occurrence of transposition-invariant matching. Finally, each character in P is first multiplied by two, and added by one.

There have been works on efficient algorithms for (δ, γ) -matching. Recent works on (δ, γ) -matching include suffix automata using bit-parallel technique [4] and fast Fourier transform [2]. For the transposition-invariant matching, several algorithms were proposed recently [5, 6, 7]. The problem with these previous approaches is that they are based on sparse dynamic programming, which is not easy to understand and implement.

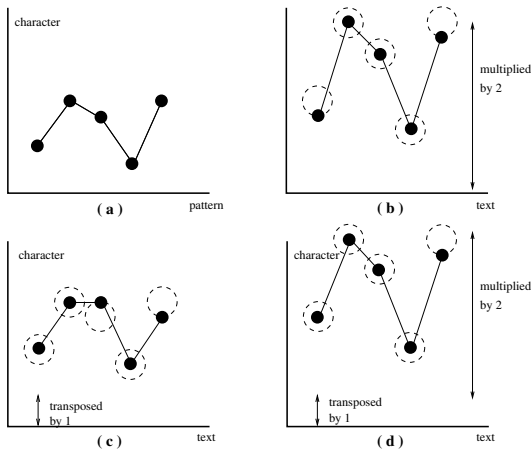


Fig. 1. In (a), the pattern is presented. (b),(c), and (d) are examples of amplified, transposition-invariant, and amplified transposition-invariant matching.

2 Preliminaries

Let T and P be strings over an integer alphabet Σ . $T[i]$ denotes the i -th character of T . $T[i, j]$ denotes the substring $T[i]T[i + 1] \cdots T[j]$. $|T| = n$ and $|P| = m$. The basic problem, (δ, γ) -matching is defined as follows.

Definition 1. Given a text $T = T[1, n]$, a pattern $P = P[1, m]$, and two parameters δ and γ , (δ, γ) -matching is to find all the substring $T[i, i + m - 1]$ such that (a) $\forall 1 \leq j \leq m, |T[i + j - 1] - P[j]| \leq \delta$ (δ -matching), and (b) $\sum_{1 \leq j \leq m} |T[i + j - 1] - P[j]| \leq \gamma$ (γ -matching).

Note that this problem makes sense if $\delta \leq \gamma \leq \delta m$. Usually δ is quite small in applications.

Clifford et al. [2] showed how to use the Fast Fourier Transform (FFT) in (δ, γ) -matching. We briefly introduce their methods. The most important property of the FFT is that all the inner-products

$$P[1, m] \cdot T[i, i + m - 1] = \sum_{j=1}^m P[j]T[i + j - 1], \quad 1 \leq i \leq n - m + 1.$$

can be calculated in $O(n \log m)$ time (see [3, Chap 32] for more details). If there is an exact match between P and $T[i, i + m - 1]$,

$$\sum_{j=1}^m (P[j] - T[i + j - 1])^2 = \sum_{j=1}^m P[j]^2 - 2P[1, m] \cdot T[i, i + m - 1] + \sum_{j=1}^m T[i + j - 1]^2$$

should be zero. The first and last terms can be computed in $O(n + m)$ time and the second term can be computed in $O(n \log m)$ time using the FFT. If there is an δ -matching between P and $T[i, i + m - 1]$, each pair of $P[j]$ and $T[i + j - 1]$ should satisfy $\prod_{\ell=-\delta}^{\delta} (P[j] - T[i + j - 1] + \ell)^2 = 0$. Hence

$$\sum_{j=1}^m \prod_{\ell=-\delta}^{\delta} (P[j] - T[i + j - 1] + \ell)^2 = 0.$$

The total time complexity for δ -matching is $O(\delta n \log m)$. For (δ, γ) -matching, they used another complex tricks to achieve the same time complexity. We do not mention more details here. For those interested, refer to [2]. Unfortunately $O(\delta n \log m)$ time for (δ, γ) -matching cannot applied to amplified matching because they require the original pattern (not the amplified one).

3 Algorithms

Now we define the problems formally and show an efficient algorithm for each problem. Our algorithms are based on Clifford et al.'s technique [2]. We first find the candidates for amplified, transposition-invariant, and amplified transposition-invariant matching. Then we verify the candidates whether they are real matches

or not. If $T[i, i + m - 1]$ is a candidate, it is easy to show that the verification takes $O(m)$ time for amplified matching and $O(\delta m)$ time for transposition-invariant and amplified transposition-invariant matching. Using the technique in [7], both verifications can be done in $O(m)$ time.

3.1 Amplified (δ, γ) -Matching

Definition 2. Given a text $T = T[1, n]$, a pattern $P = P[1, m]$, and two parameters δ and γ , the amplified (δ, γ) -matching is to find all the substring $T[i+m-1]$ such that for an integer α , (a) $\forall 1 \leq j \leq m, |T[i+j-1] - \alpha \times P[j]| \leq \delta$, and (b) $\sum_{1 \leq j \leq m} |T[i+j-1] - \alpha \times P[j]| \leq \gamma$.

The difference lies in the fact that we need to determine α . Once α is known, $\sum_{j=1}^m (\alpha \times P[j] - T[i+j-1])^2 = \alpha^2 \times \sum_{j=1}^m P[j]^2 - 2\alpha \times P[1, m] \cdot T[i, i+m-1] + \sum_{j=1}^m T[i+j-1]^2$. And for δ -matching,

$$\sum_{j=1}^m \prod_{\ell=-\delta}^{\delta} (\alpha \times P[j] - T[i+j-1] + \ell)^2 = 0$$

can be computed in $O(\delta n \log m)$ time.

To store α , we use an integer array $\alpha[1, m]$. First we select a base element $P[k]$. For simplicity, assume that $P[k]$ is the greatest in P . Then $\alpha[i] = \lfloor T[i+k-1]/P[k] \rfloor, 1 \leq i \leq n-k+1$. While computing the FFT to find a match between $T[i, i+m-1]$ and P , we use $\alpha[i]$.

The base element $P[k]$ should meet one condition. For any character $T[i]$, $\alpha[i] - 0.5 \leq T[i]/P[k] < \alpha[i] + 0.5$. If there is an occurrence of (δ, γ) -matching at position i , $|T[i] - \alpha[i] \times P[k]| \leq \delta$. By replacing $T[i]$ with $\alpha[i] \times P[k] - \delta$ and $\alpha[i] \times P[k] + \delta$, we get two inequalities,

$$\alpha[i] - 0.5 \leq \frac{\alpha[i] \times P[k] - \delta}{P[k]} < \alpha[i] + 0.5 \text{ and}$$

$$\alpha[i] - 0.5 \leq \frac{\alpha[i] \times P[k] + \delta}{P[k]} < \alpha[i] + 0.5.$$

After some tedious computation using the fact $\delta \geq 0$, we get $P[k] > 2\delta$. Note that it doesn't mean that all the characters in P should meet this condition. Just one character is enough. Fortunately this condition can be met easily in applications. For example, A=440Hz in music analysis and human being cannot hear lower than 20Hz. Hence we can use δ up to 10, which is enough. Note that the answers from the FFT is just considering δ -matching. For (δ, γ) -matching, we need the verification mentioned above.

Theorem 1. The amplified (δ, γ) -matching can be solved in $O(\delta n \log m + occ \times m)$ time, where occ is the number of candidates.

Proof. Computing the array D takes $O(n)$ time. The FFT runs in $O(\delta n \log m)$ time. After finding occ candidates, each requires $O(m)$ time verification.

3.2 Transposition-Invariant Matching

Definition 3. Given a text $T = T[1, n]$, a pattern $P = P[1, m]$, and two parameters δ and γ , the transposition-invariant (δ, γ) -matching is to find all the substring $T[i, i + m - 1]$ such that for an integer β , (a) $\forall 1 \leq j \leq m$, $|T[i + j - 1] - (P[j] + \beta)| \leq \delta$, and (b) $\sum_{1 \leq j \leq m} |T[i + j - 1] - (P[j] + \beta)| \leq \gamma$.

Instead of using sparse-dynamic programming, we use a simpler method. We create two new strings $T' = T'[1, n - 1]$ and $P' = P'[1, m - 1]$ such that $T'[i] = T[i + 1] - T[i]$ and $P'[i] = P[i + 1] - P[i]$. Then the following simple lemma holds.

Lemma 1. If there is a (δ, γ) -matching of P at position i of T , then there is a $(2\delta, 2\gamma)$ -matching of P' at position i of T' .

Proof. We first begin proving 2δ -matching. If there is an occurrence of δ -matching at position i of T , it is evident that $-\delta \leq T[i + j - 1] - P[j] \leq \delta$ and $-\delta \leq T[i + j] - P[j + 1] \leq \delta$ for $1 \leq j \leq m - 1$. From these equations, it follows that $-2\delta \leq (T[i + j] - T[i + j - 1]) - (P[j + 1] - P[j]) \leq \delta$. Since $T'[i + j - 1] = T[i + j] - T[i + j - 1]$ and $P'[j] = P[j + 1] - P[j]$, $-2\delta \leq T'[i + j - 1] - P'[j] \leq 2\delta$. Now we prove 2γ -matching. Now we consider about gamma-matching. If there is γ -matching of P in T at position i , it should be $\sum_{j=1}^m |T[i + j - 1] - P[j]| \leq \gamma$. Now we use a simple fact $|A| + |B| \geq |A + B|$.

$$\begin{aligned} \sum_{j=1}^{m-1} |T'[i + j - 1] - P'[j]| &= \sum_{j=1}^{m-1} |(T[i + j] - T[i + j - 1]) - (P[i + 1] - P[i])| \\ &= \sum_{j=1}^{m-1} |(T[i + j] - P[i + 1]) + (P[i] - T[i + j - 1])| \\ &\leq \sum_{j=1}^{m-1} (|(T[i + j] - P[i + 1])| + |P[i] - T[i + j - 1]|) \\ &\leq 2 \times \sum_{i=1}^m |T[i + j - 1] - P[i]| \\ &\leq 2\gamma. \end{aligned}$$

Using this fact, we find occurrences of $(2\delta, 2\gamma)$ -matching of P' from T' . The results are candidates for (δ, γ) -matching of P from T . Then we check whether they are real occurrences of (δ, γ) -matching or not.

Theorem 2. The transposition-invariant (δ, γ) -matching can be solved in $O(\delta n \log m + occ \times m)$ time, where occ is the number of candidates.

Proof. Computing T' and P' takes in $O(m + n)$ time. The FFT runs in $O(\delta n \log m)$ time. Each verification runs in $O(m)$ time.

Now we consider how large occ is. If T and P are drawn randomly from Σ , it is easy to show that the probability that $T'[i]$ and $P'[j]$ can have a 2δ -matching is $(4\delta + 1)/|\Sigma|$. Hence, the probability is $((4\delta + 1)/|\Sigma|)^{m-1}$. The expected number of candidates is $n((4\delta + 1)/|\Sigma|)^{m-1}$, which is small when δ is quite small and $|\Sigma|$ is large.

3.3 Amplified Transposition-Invariant Matching

We simply explain the outline of amplified transposition-invariant matching. The first observation is that if there is an occurrence of amplified transposition-invariant matching of P in T , then there is an occurrence of amplified matching of P' in T' (we can get P' and T' as we did in transposition-invariant matching). Therefore, we first create P' and T' , then we do amplified matching. Then we verify the results as we did in transposition-invariant matching, using α array obtained during amplified matching. It is easy to show that the time complexity is $O(\delta n \log m + occ \times m)$.

4 Conclusion

We showed simple $O(\delta n \log m + occ \times m)$ time algorithms for amplified, transposition-invariant, and amplified transposition-invariant matching. Further research includes parameterised version of the problems discussed in this paper, which means finding a mapping $\Sigma \rightarrow \Sigma'$. For exact scaled matching, there is an algorithm for the parameterised version [1]. Another interesting problem is to find occurrences of (δ, γ) -matching by more complex transforms.

References

1. A. Amir, A. Butman, and M. Lewenstein. Real scaled matching. *Information Processing Letters*, 70(4):185–190, 1999.
2. P. Clifford, R. Clifford, and C. S. Iliopoulos. Faster Algorithms for (δ, γ) -Matching and Related Problems. In *Proc. of 16th Combinatorial Pattern Matching (CPM '05)*, pages 68–78. 2005.
3. T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. MIT Press, 1990.
4. M. Crochemore, C. S. Iliopoulos, G. Navarro, Y. Pinzón, and A. Salinger. Bit-parallel (δ, γ) -matching suffix automata. *Journal of Discrete Algorithms*, 3(2-4):198-214, 2004.
5. H. Hyrrö. Restricted Transposition Invariant Approximate String Matching. In *Proc. of 12th String Processing and Information Retrieval (SPIRE '05)*, pages 257–267, 2005.
6. K. Lemström, G. Navarro, and Y. Pinzón. Practical Algorithms for Transposition-Invariant String-Matching. *Journal of Discrete Algorithms*, 3(2-4):267-292, 2005
7. V. Mäkinen, G. Navarro, and E. Ukkonen. Transposition Invariant String Matching. *Journal of Algorithms*, 56(2):124-153, 2005.

SOM and Neural Gas as Graduated Nonconvexity Algorithms*

Ana I. González, Alicia D'Anjou, M. Teresa García-Sebastian, and Manuel Graña

Grupo de Inteligencia Computacional,
Facultad de Informática, UPV/EHU,
Apdo. 649, 20080 San Sebastián, España
ccpgrrom@si.ehu.es

Abstract. Convergence of the Self-Organizing Map (SOM) and Neural Gas (NG) is usually contemplated from the point of view of stochastic gradient descent. This class of algorithms is characterized by a very slow convergence rate. However we have found empirically that One-Pass realizations of SOM and NG provide good results or even improve over the slower realizations, when the performance measure is the distortion. One-Pass realizations use each data sample item only once, imposing a very fast reduction of the learning parameters that does not conform to the convergence requirements of stochastic gradient descent. That empirical evidence leads us to propose that the appropriate setting for the convergence analysis of SOM, NG and similar competitive clustering algorithms is the field of Graduated Nonconvexity algorithms. We show they can easily be put in this framework.

1 Introduction

In this paper we focus on two well known competitive artificial neural networks: Self Organising Map (SOM) [3, 14, 15] and the Neural Gas (NG) [17]. They can be used for Vector Quantization [1, 9, 10, 12], which is a technique that maps a set of input vectors into a finite collection of predetermined codevectors. The set of all codevectors is called the *codebook*. In designing a vector quantizer, the goal is to construct a codebook for which the expected distortion of approximating any input vector by a codevector is minimized. Therefore, the distortion computed over the sample data is the natural performance feature for VQ algorithms. It is important to note this, because for the SOM there is a body of work devoted to its convergence to organized states (i.e.[22]), however for the NG there is no meaning for such a concept. Organized states are related to the nonlinear dimension reduction ability of the SOM based on topological preservation properties of the organized states. The starting assumption in this paper is that both SOM and NG are to be used as VQ design algorithms. Some works [3] have already pointed that SOM may be viewed as a robust initialization step for the fine-tuning of the Simple Competitive Learning (SCL) to perform VQ.

Both SOM and NG algorithms have the appearance of stochastic gradient descent (online) algorithms [8] in their original definitions, that is, whenever an input vector is

* The work is partially supported by MEC grants DPI2003-06972 and VIMS-2003-20088-c04-04, and UPV/EHU grant UE03A07.

presented, a learning (adaptation) step occurs. It has been shown that an online version of the NG algorithm can find better local solutions than the online SOM [17]. Online realizations are very lengthy due to the slow convergence rate of the stochastic gradient descent. To speed up computations, there are batch versions for both algorithms. Batch realizations correspond to deterministic gradient descent algorithms. The parameter estimation is performed using statistics computed over the whole data sample. The batch version of SOM was already proposed in [14] as a reasonable speed-up of the online SOM, with minor solution quality degradation. In the empirical analysis reported in [6], the main drawbacks for the batch SOM are its sensitivity to initial conditions and the bad organization of the final class representatives that may be due to poor topological preservation. Good initialisation instances of the batch SOM may improve the solutions obtained by the online SOM. On the other hand, the online SOM is robust against bad initializations and provides good topological ordering, if the adaptation schedule is smooth enough. The batch version of the NG algorithm has been studied in [18] as an algorithm for clustering data. It has been proposed as a convenient speed-up of the online NG.

Both the online and batch algorithm versions imply the iteration over the whole sample several times. On the contrary, One-Pass realizations visit only once the sample data. This adaptation framework is not very common in the neural networks literature; in fact, the only related reference that we have found is [4]. The effective scheduled sequences of the learning parameters applied to meet the fast adaptation requirement fall far from the theoretical convergence conditions. However, as we shall see, in some cases the distortion results are competitive with the conventional SOM and NG online and batch versions. If we take into account the computation time, the One-Pass realization superiority becomes spectacular. These results lead us to think that may be there are more critical phenomena working in the convergence other than the learning rate. We postulate that both SOM and NG are instances of the Graduated Nonconvexity (GNC) algorithms [19, 20], which are related to the parameter continuation methods [2]. GNC algorithms try to solve the minimization of a non-convex objective function by the progressive search of the minima of a sequence of functions that depending of a parameter are morphed from a convex function up to the non-convex original function. Continuation methods perform the search for roots of highly nonlinear systems in a similar way.

For independent verification, the Matlab code of the experiments described in this paper is available in the following web address: www.sc.ehu.es/acwgoaca/ under the heading “proyectos”.

Section 2 presents the formal definition of the algorithms. Section 3 gives the experimental results. Section 4 discusses the formulation of the SOM and NG as GNC algorithms. Section 5 is devoted to conclusions and discussion.

2 Algorithm Definitions

Let it be $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ the input data sample real valued vectors and $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_c\}$ the set of real valued codevectors (*codebook*). The design of the codebook is performed minimizing the error/distortion function E :

$$E = \sum_{i=1}^n \left\| \mathbf{x}_i - \mathbf{y}_{k(i)} \right\|^2; \quad k(i) = \underset{j=1, \dots, c}{\operatorname{argmin}} \left\{ \left\| \mathbf{x}_i - \mathbf{y}_j \right\|^2 \right\} \quad (1)$$

Each algorithm described below has some control parameters, like the learning ratio, the neighbourhood size and shape, or the temperature. The online realizations usually modify their values following each input data presentation and adaptation of the codebook. The batch realizations modify their values after each presentation of the whole input data sample. Both online and batch realizations imply that the input data set is presented several times. On the contrary, the One-Pass realizations imply that each input data is presented at most once for adaptation, and that the control parameters are modified after each presentation.

2.1 One-Pass Version of Self-Organizing Map and Neural Gas

The SOM is a particular case of the general Competitive Neural Network algorithm:

$$\mathbf{y}_i(t+1) = \mathbf{y}_i(t) + \alpha(t) H_i(\mathbf{x}(t), \mathbf{Y}(t)) (\mathbf{x}(t) - \mathbf{y}_i(t)) \quad (2)$$

Where t is the order of presentation of sample vectors; the size of the sample is n . We denote by $H_i(\mathbf{x}, \mathbf{Y})$ the so-called neighbouring function, and by $\alpha_i(t)$ the (local) learning rate. In the case of a conventional online realization, t corresponds to the iteration number over the sample, and the learning rate and neighbour value is fixed during iteration. In the case of One-Pass realization, t corresponds to the input vector presentation number, and the learning rate and neighbour value is updated during iteration. In their general statement, Competitive Neural Networks are designed to perform stochastic gradient minimisation of a distortion-like function similar to that in equation (1). In order to guarantee theoretical convergence, the learning rate must comply with the following conditions:

$$\lim_{t \rightarrow \infty} \alpha(t) = 0, \sum_{t=0}^{\infty} \alpha(t) = \infty, \sum_{t=0}^{\infty} \alpha^2(t) < \infty \quad (3)$$

However, these conditions imply very lengthy adaptation processes, for finite samples they usually force walking over the sample several times. The idea of performing a One-Pass realization of the minimization process violates these conditions. Besides, it imposes strong schedules of the algorithm control parameters. In the experiments, the learning rate follows the expression [5]:

$$\alpha(t) = \alpha_0 \left(\alpha_n / \alpha_0 \right)^{\frac{t}{n}} \quad (4)$$

Where α_0 and α_n are the initial and final value of the learning rate, respectively. Therefore after n presentations the learning rate reaches its final value. In the case of the SOM, the neighbouring function is defined over the space of the neuron (codevector) indices. In the experiments reported in this paper, we assume a 1D topology of the codevector indices. The neighbourhoods considered decay exponentially following the expression:

$$H_i(\mathbf{x}, \mathbf{Y}) = \begin{cases} 1 & |w - i| \leq \left\lfloor h_0(h_n/h_0)^{8t/n} \right\rfloor \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$w = \operatorname{argmin}_{\{k=1, \dots, c\}} \|\mathbf{x} - \mathbf{y}_k\|^2 \quad 1 \leq i \leq c$$

The initial and final neighbourhood radius are h_0 and h_n , respectively. The expression ensures that the neighbouring function reduces to the simple competitive case (null neighbourhood) after the presentation of the first $1/8$ inputs of the sample. With this neighbourhood reduction rate, we obtain, after a quick initial ordering of the codevectors, a slow local fine-tuning. We proposed this scheduling in [11] to approach real-time constraints and other authors have worked with this idea [3] in the context of conventional online realizations.

The NG introduced in [17] shares with the SOM the structure shown in equation (2) it is characterized by the following neighbouring function:

$$H_i(\mathbf{x}, \mathbf{Y}) = \exp(-\operatorname{ranking}(i, \mathbf{x}, \mathbf{Y})/\lambda) \quad (6)$$

The *ranking* function returns the position $\{0, \dots, c - 1\}$ of the codevector \mathbf{y}_i in the set of codevectors ordered by their distances to the input \mathbf{x} . All codevectors are updated, there are not properly defined neighbours, but the temperature parameter λ decays exponentially according to the following expression:

$$\lambda(t) = \lambda_0(\lambda_n/\lambda_0)^{\frac{t}{n}} \quad (7)$$

Where λ_0 and λ_n are its initial and final value. The expression ensures that the neighbouring function reduces to the simple competitive case (null neighbourhood) as happens with SOM.

In the case of his online version, t would correspond to the input vector presentation number, and the temperature parameter value would be fixed for all the input samples during a complete presentation of the input data set.

2.2 Batch Version of Self-Organizing Map and Neural Gas

Kononen’s Batch Map [14, 15] defined the batch version of the SOM algorithm. Among its advantages, there is no learning rate parameter and the computation is faster than the conventional online realization. This algorithm can be viewed as the LBG algorithm [16] plus a neighbouring function. When the input data set is completely presented, each input sample is classified in the Voronoi region defined by the winner codevector $y_w(t)$:

$$\forall i, 1 \leq i \leq c, \|x(t) - y_w(t)\| \leq \|x(t) - y_i(t)\| \quad (8)$$

Where t corresponds to the iteration number over the sample and learning parameter values are fixed during each iteration. To recalculate the codebook, each codevector is computed as the centroid of its Voronoi region and the Voronoi regions of his neighbour codevectors as follows:

$$y_i(t) = \sum_{x(t) \in U_i} x(t) / n(U_i) \quad (9)$$

Where U_i is the union of Voronoi regions corresponding to the codevectors that lie up to a certain radius $h(t)$ from codevector i , in the topology of the codevector indices. And $n(U_i)$ means the number of samples $x(t)$ that belong to U_i . To determine the radius of the neighbourhood we applied the following function:

$$h(t) = \left[h_0 (h_n / h_0)^{\frac{t}{n}} \right] - 1 \quad (10)$$

The expression ensures that the neighbouring function reduces to h_n after n iterations. In the experiments, it takes value 0.1, which implies that its operation is equivalent to LBG or the k-means: the codevector is the arithmetic mean of the input sample that lies in the unique Voronoi region associated with the codevector.

In the Batch SOM, all neighbours have the same contribution to the centroid calculation, as in the SOM online realization. A definition of Batch NG arises from the idea of changing the contribution to the codebook in function of neighbour-region distances, imitating the online realization of NG. We produce this effect applying a weighting mean as follows:

$$y_i(t) = \sum_{x(t)} x(t) w_{x(t)} / \sum w_{x(t)} \quad (11)$$

Where $w_{x(t)}$ is the weighting term for the input samples in the Voronoi region, defined by the codevector y_i , given by the following expression:

$$w_{x(t)} = \exp(-\text{ranking}(i, \mathbf{x}(t), \mathbf{Y}(t)) / \lambda(t)) \quad (12)$$

The ranking function and temperature parameter λ are equal to those in the One-Pass case. The neighbour-region contribution decays exponentially due to the evolution of λ in equation (6). As with the Batch SOM, the Batch NG converges to the LBG algorithm: only the region corresponding to the codebook contributes to its calculus.

3 Experimental Results

The results presented here are continuation of the ones reported in [21], we have applied the algorithms to higher dimension data. The figure 1 shows two 3D data sets with some SOM VQ solution. The 2D versions have been used in [5,7] for evaluation of clustering and VQ algorithms. Figure 1(a) is three-level Cantor set distribution on 3D space. Figure 1(b) is a mixture of Gaussians in 3D space.

The three-level Cantor set (2048 data points) is uniformly distributed on a fractal; it is constructed by starting with a unit interval, removing the middle third, and then recursively repeating the procedure on the two portions of the interval that are left. And the third data set is a collection of data points generated by a mixture of ten Gaussian distributions.

The codebook initialization used in this paper and reflected in the results of figure 2 is a random selection of input sample data. The codebook size is set to $c = 16$

codevectors. The maximum number of sample presentations has been established at $n = 50$ for conventional online and batch realizations of the algorithms. Nevertheless, we introduce a stopping criterion on the relative decrement of the distortion; the process will stop if it is not greater than $\xi = 0.001$.

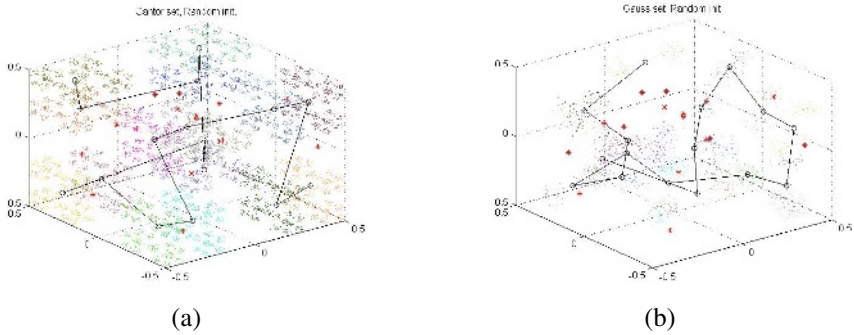


Fig. 1. The 3D benchmark data sets (a) Cantor set and (b) mixture of Gaussian distributions, with sample VQ solutions obtained by the SOM

For SOM algorithms the neighbourhood initial and final parameter values have been set to: $h_0 = c/2 + 1$; $h_n = 0.1$, and for NG algorithms they have been set to $\lambda_0 = c/2$; $\lambda_n = 0.01$. In both One-Pass version algorithms the learning rate values are $\alpha_0 = 0.5$ and $\alpha_n = 0.005$. We have executed 100 times each algorithm. In figures 2a, 2b the y-axis corresponds to the mean and 0.99 confidence interval of the product of the final distortion and the computation time as measured by Matlab. The algorithms tested are: the online conventional realizations of SOM and NG, the batch versions (BSOM and BNG) and the online One-Pass realizations (SOMOP and NGOP).

The inspection of the figure 2 reveals that the relative efficiencies of the algorithms measured by the distortion depend on the nature of the data. From our point of view, the most salient feature of the distortion plots is that the One-Pass realizations are competitive with the batch and online realizations. Although this is not the main concern of this paper, the distortion results show that the NG improves the SOM most of the times, confirming the results in the literature [17]. The batch realization sometimes improves the online realization (Cantor), sometimes not (Gaussian).

When we take into account the computation time in the plots of figures 2a, 2b, the improvement of the One-Pass realization over the batch and online conventional realizations is spectacular. It can also be appreciated the improvement of the batch realization over the conventional online realization. We have used the product of time and distortion instead of the ratio distortion/time in order to maintain the qualitative interpretation of the plots: the smallest are the best. The figure 3 shows the evolution of the error for some instances of the algorithms. It can be appreciated that the One-Pass realizations achieve really fast the error values equivalent to those of the conventional realizations.

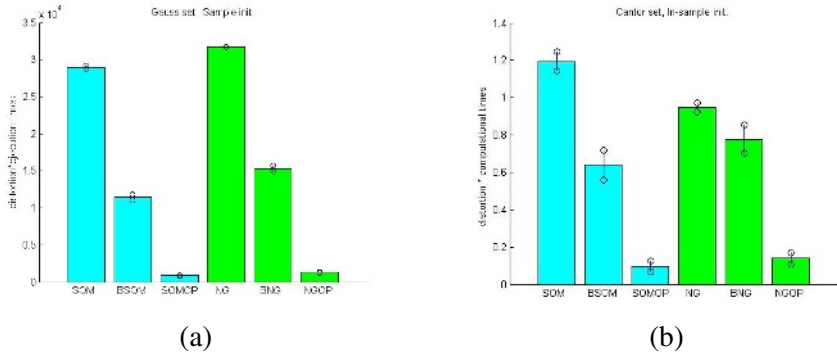


Fig. 2. Results on the three benchmark algorithms. The product of distortion and the computational time for (a) Gaussian data, (b) cantor data.

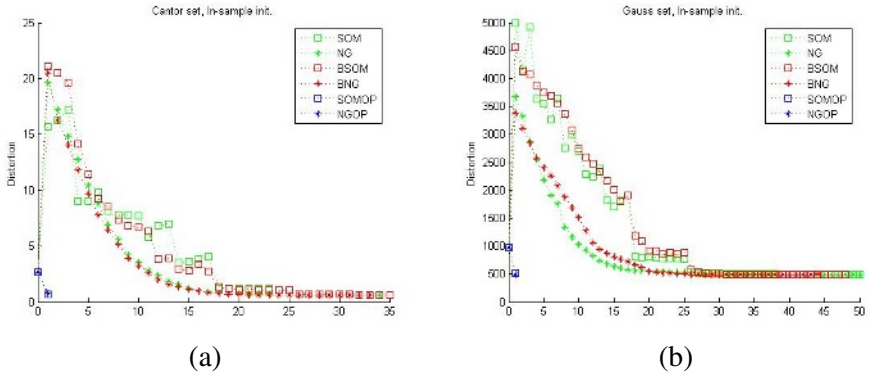


Fig. 3. Sample plots of the evolution of the quantization distortion along the training process over the (a) Cantor set, (b) mixture of Gaussian data.

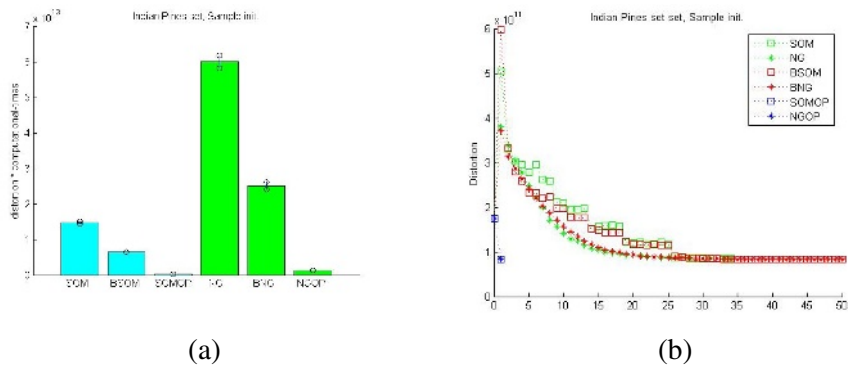


Fig. 4. (a) the product of the distortion and the computation time for the Indian Pines image data, (b) samples of the error trajectories during training

The last experimental data is an AVIRIS hyperspectral image, the so called Indian Pines image, used and described, for instance in [23]. The image is a 145x145 pixel image, where each pixel has 220 bands. The figure 4 shows the average efficiency of the algorithms realizations and a sample trajectory of the error during the training process. Again, in this high dimensional data, the One-Pass realization is much more efficient and gives distortion results comparable with the Batch and conventional online realizations.

4 SOM and NG as GNC

The previous results cannot be understood in the framework of the convergence of the stochastic gradient algorithms. The basic formulation of the GNC approach [19, 20] is that the function to be minimized is the MAP estimate of a sampled surface corrupted by additive noise $M(x)=D(x)+N(x)$. This MAP estimate $p(R=D|M)$ takes the form

$$E[R]=-\log p(M|R=D)-\log(D=R)=E_d[R]+E_s[R], \tag{13}$$

where $E_d[R]$ is the data term and $E_s[R]$ is the smoothness term. The data term is quadratic under the usual Gaussian noise assumption, and the smoothness term express any *a priori* information about the surface. In [20] the smoothness term is formulated over the surface gradient. The GNC function general formulation is:

$$E[R]=\sum_x (M(x)-R(x))^2 + E_s[R], \tag{14}$$

where the smoothness term depends on some parameter $E_s[R]=f_\sigma(R)$. The key of GNC methods is that the function to be minimized $E[R]$ is embedded in a one-parameter functional family $E_\sigma[R]$ so that the initial functional $E_{\sigma_0}[R]$ is convex, and the final functional is equivalent to the original function $E_0[R]\equiv E[R]$. The minimization is performed tracking the local minimum of $E_\sigma[R]$ from the initial to the final functional. As the initial functional is convex, the algorithm becomes independent of the initial conditions. It must be noted that one of the properties that the SOM and NG show over the bare SCL algorithms is the robustness against bad initial conditions [3].

The NG was proposed [17] as the minimization of the following functional,

$$E_{ng}(w,\lambda)=\frac{1}{2C(\lambda)}\sum_{i=1}^N \int d^D v P(v) h_\lambda(k_i(v,w))(v-w_i)^2, \tag{15}$$

that we discretize here, assuming sample data $\{v_1, v_2, \dots, v_M\}$,

$$E_{ng}(w,\lambda)=\frac{1}{2C(\lambda)}\sum_{i=1}^N \sum_{j=1}^M h_\lambda(k_i(v_j,w))(v_j-w_i)^2, \tag{16}$$

where $k_i(v_j, w)$ is the ranking function. Note that we can reorganize it as follows:

$$E_{ng}(w, \lambda) = \sum_{i=1}^N h_\lambda(k_i(v_i, w))(v_i - w_i)^2 + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^M h_\lambda(k_i(v_j, w))(v_j - w_i)^2 \tag{17}$$

if $h_\lambda(k_i(v_i, w))$ is constant and equal to one, the first term in equation (17) is equivalent to the data term in equation (14). This may be so if there is a specific ordering of the weights so that they match the first N data points. If, additionally, the value of $h_\lambda(k_i(v_i, w))$ is one for the best matching unit, then the coincidence is exact. In the NG this function is an exponential function that performs a role like the focusing Gaussians in [20]. The second term in equation (17) corresponds to the smoothing term in equation (14). One key problem in GNC is to ensure that the initial functional is convex [20]. Other problem is to ensure that there are no bifurcations or other effects that may affect the continuation process. It seems that for NG and SOM it is very easy to ensure the convexity of the initial functional and that the continuation is also an easy process. In the case of the SOM, when the neighbourhood function is the one used in the experiments it is assumed that the functional to be minimized is the extended distortion:

$$E_{SOM}(\mathbf{Y}, \lambda) = \sum_{i=1}^N \sum_{j=1}^M H_i(\mathbf{x}_j, \mathbf{Y})(\mathbf{x}_j - \mathbf{y}_i)^2, \tag{18}$$

$$H_i(\mathbf{x}, \mathbf{Y}) = \begin{cases} 1 & |w - i| \leq \lambda \\ 0 & \text{otherwise} \end{cases}; \quad w = \operatorname{argmin}_{\{k = 1, \dots, c\}} \|\mathbf{x} - \mathbf{y}_k\|^2, 1 \leq i \leq c$$

Again it is easy to decompose the functional in a structure similar to that of equation (14).

$$E_{SOM}(\mathbf{Y}, \lambda) = \sum_{j=1}^M (\mathbf{x}_j - \mathbf{y}_w)^2 + \sum_{i=1}^N \sum_{\substack{j=1 \\ i \neq w}}^M H_i(\mathbf{x}_j, \mathbf{Y})(\mathbf{x}_j - \mathbf{y}_i)^2, \tag{19}$$

Therefore, the SOM can be assimilated to a GNC algorithm. It is very easy to ensure the convexity of its functional for large neighbourhoods, and the continuation process seems to be very robust.

5 Discussion and Conclusions

The paradoxical results found in [23] and extended here, showing that the One-Pass realization of the SOM and NG can give competitive performance in terms of distortion, and much better than the “conventional” batch and online realizations in terms of computational efficiency (time x distortion) leads us to the idea that these algorithm’s performance is more sensitive to the neighbourhood parameters than to the learning gain parameter. We think that the context of continuation methods [2] and GNC [19,20] is an adequate context to analyze the convergence properties of the

algorithms. We have shown that the SOM and NG minimized functional can be easily seen to be like the ones considered in GNC. We will devote our future efforts to deepen in the analysis of these algorithms from this point of view.

References

- [1] S.C. Ahalt, A.K. Krishnamurthy, P. Chen, D.E. Melton (1990), Competitive Learning Algorithms for Vector Quantization, *Neural Networks*, vol. 3, p.277-290.
- [2] E. L. Allgower and K. Georg, Numerical Continuation Methods. An Introduction, Vol. 13, Springer Series in Computational Mathematics. Berlin/Heidelberg, Germany: Springer-Verlag, 1990.
- [3] E. Bodt, M. Cottrell, P. Letremy, M. Verleysen (2004), On the use of self-organizing maps to accelerate vector quantization, *Neurocomputing*, vol 56, p. 187-203.
- [4] C. Chan, M. Vetterli (1995), Lossy Compression of Individual Signals Based on String Matching and One-Pass Codebook Design, *ICASSP'95*, Detroit, MI.
- [5] C. Chinrungrueng, C. Séquin (1995), Optimal Adaptive K-Means Algorithm with Dynamic Adjustment of Learning Rate, *IEEE Trans. on Neural Networks*, vol. 6(1), p.157-169.
- [6] Fort J.C., Letrémy P., Cottrell M. (2002), Advantages and Drawbacks of the Batch Kohonen Algorithm, in M. Verleysen (ed), Proc. of ESANN'2002,Brugge, Editions D Facto, Bruxelles, p. 223-230.
- [7] B. Fritzke (1997), The LBG-U method for vector quantization - an improvement over LBG inspired from neural networks, *Neural Processing Letters*, vol. 5(1) p. 35-45.
- [8] K. Fukunaga (1990), *Statistical Pattern Recognition*, Academic Press.
- [9] A. Gersho (1982), On the structure of vector quantizers, *IEEE Trans. Inf. Th.*, 28(2), p.157-166.
- [10] A. Gersho, R.M. Gray (1992), *Vector Quantization and signal compression*, Kluwer.
- [11] A. I. Gonzalez, M. Graña, A. d'Anjou, F.X. Albizuri (1997), A near real-time evolutive strategy for adaptive Color Quantization of image sequences, *Joint Conference Information Sciences*, vol. 1, p. 69-72.
- [12] R.M. Gray (1984), Vector Quantization, *IEEE ASSP*, vol. 1, p.4-29.
- [13] T. Hofmann and J. M. Buhmann (1998) Competitive Learning Algorithms for Robust Vector Quantization *IEEE Trans. Signal Processing* 46(6): 1665-1675
- [14] T. Kohonen (1984) (1988 2nd ed.), *Self-Organization and associative memory*, Springer Verlag.
- [15] T. Kohonen (1998), The self-organising map, *Neurocomputing*, vol 21, p. 1-6.
- [16] Y. Linde, A. Buzo, R.M. Gray (1980), An algorithm for vector quantizer design, *IEEE Trans. Comm.*, 28, p.84-95.
- [17] T. Martinez, S. Berkovich, K. Schulten (1993), Neural-Gas network for vector quantization and his application to time series prediction, *IEEE trans. Neural Networks*, vol. 4(4), p.558-569.
- [18] S. Zhong, J. Ghosh (2003) A Unified Framework for Model-based Clustering *Journal of Machine Learning Research* 4:1001-1037
- [19] A. Blake, and A. Zisserman, Visual Reconstruction. Cambridge, Mass.: MIT Press, 1987.
- [20] Nielsen, M. (1997) Graduated nonconvexity by functional focusing, *IEEE Trans. Patt. Anal. Mach. Int.* 19(5):521 – 52
- [21] A.I. Gonzalez, M. Graña (2005) Controversial empirical results on batch versus one pass online algorithms, Proc. WSOM2005, Sept. Paris, Fr. pp.405-411
- [22] J.C. Fort, G. Pagès (1996) About the Kohonen algorithm: strong or weak Self-organization? *Neural Networks* 9(5) pp.773-785
- [23] Gualtieri, J.A.; Chettri, S. 'Support vector machines for classification of hyperspectral data', Proc. Geosci. Rem. Sens. Symp., 2000, IGARSS 2000. pp.:813 - 815 vol.2

Analysis of Multi-domain Complex Simulation Studies

James R. Gattiker, Earl Lawrence, and David Higdon

Los Alamos National Laboratory
{gatt, earl, dhigdon}@lanl.gov

Abstract. Complex simulations are increasingly important in systems analysis and design. In some cases simulations can be exhaustively validated against experiment and taken to be implicitly accurate. However, in domains where only limited validation of the simulations can be performed, implications of simulation studies have historically been qualitative. Validation is notably difficult in cases where experiments are expensive or otherwise prohibitive, where experimental effects are difficult to measure, and where models are thought to have unaccounted systematic error. This paper describes an approach to integrate simulation experiments with empirical data that has been applied successfully in a number of domains. This methodology generates coherent estimates of confidence in model predictions, model parameters, and estimates, i.e. calibrations, for unobserved variables. Extensions are described to integrate the results of separate experiments into a single estimate for simulation parameters, which demonstrates a new approach to model-based data fusion.

1 Introduction

Computational simulation applications are increasingly used to explore a number of domains, including: climate, ocean, and weather modeling; atomic scale physics modeling; aerodynamic modeling; and cosmology applications. A significant challenge for using simulation studies is the quantitative analysis of simulation results, and the comparison and integration of the simulations with experimental data.

At Los Alamos National Laboratory, a challenge is to certify the safety and reliability of nuclear weapons where only indirect physical experiments can be performed[1]. Simulations model physical experimental results. Uncertainties arise from a variety of sources that include: uncertainty in the specification of initial conditions, uncertainty in the value of important physical constants (e.g., melting temperatures, equations of state, stress-strain relationships, shock propagation, and transient states), inadequate mathematical models, and numerical computation effects. Experimental observations constrain uncertainties within the simulator, and are used to validate simulation components and responses[10].

The methodology described here addresses three main goals in simulation analysis. First is the *quantification of uncertainty* in predictions. Most simulations systems lack the ability to directly assess the uncertainty in their results,

although it is clear from failure to match reality that both bias and uncertainty exist. The second goal is the *calibration of unknown parameters*. Simulations often have parameters that are either non-physical or are unmeasurable in experiments, and must be determined, or calibrated. The third goal addressed, discussed here for the first time, is the linking and *joint calibration* of variables common to separate experiments.

Additional issues constrain approaches to this problem. Experimental data in typical application areas is generally difficult to collect because of expense, difficulty in making physical measurements, or external constraints. Simulation studies are often computationally expensive, having usually been developed at the limits of feasible computation, and so there is limited access to alternative simulation parameter settings. Some exploration of alternatives is therefore possible, but putting the simulator directly into an iterated method can be prohibitive.

This paper describes a methodology that has been implemented and demonstrated to be effective in addressing these issues in real problems[2]. The approach is to model simulation response with an accurate emulated response. Parameters of this emulator as well as simulation parameters are simultaneously determined using a Bayesian parameter formulation and associated Markov chain Monte Carlo sampling. The emulator itself is a stochastic process model, modeling both simulation response and systematic bias.

1.1 The Model Evaluation Problem

This section follows an explanatory “toy” problem that captures many of the issues in analyzing (simulation) models in conjunction with experiments[3]. The task is to analyze models of gravitational attraction, through the combination of an analytical model and experiment. To study the nature of falling objects a test object is dropped from various heights to study their descent time. In addition to characterizing the measurements of limited experiments, we wish to extrapolate the behavior. In practice, the drop time is not a simple effect due to atmospheric effects. For explanatory purposes, experimental data is generated according to

$$\frac{d^2 z}{dt^2} = -1 - 0.3 \frac{dz}{dt} + \epsilon,$$

which includes the square law of gravitational attraction, plus a term for linear effects. If the simulation models only the gravitational attraction $\frac{d^2 z}{dt^2} = -1$, the results do not explain the data well, and extrapolate even more poorly. This model would give a fitted response as shown in Fig. 1a.

To analyze the simulation results, we need a model of the data that explicitly allows for systematic error in the simulation system. In this approach, we use a model η that models the simulation responses, and an additional model δ of the discrepancy between the simulation and the data, so that the model is comprehensive, i.e., $Y(z) = \eta(z) + \delta(z) + \epsilon$. The data can now be modeled accurately as a systematic discrepancy from the simulation. Incorporating uncertainty into the model response, the results are shown in Fig. 1b. The model’s η ,

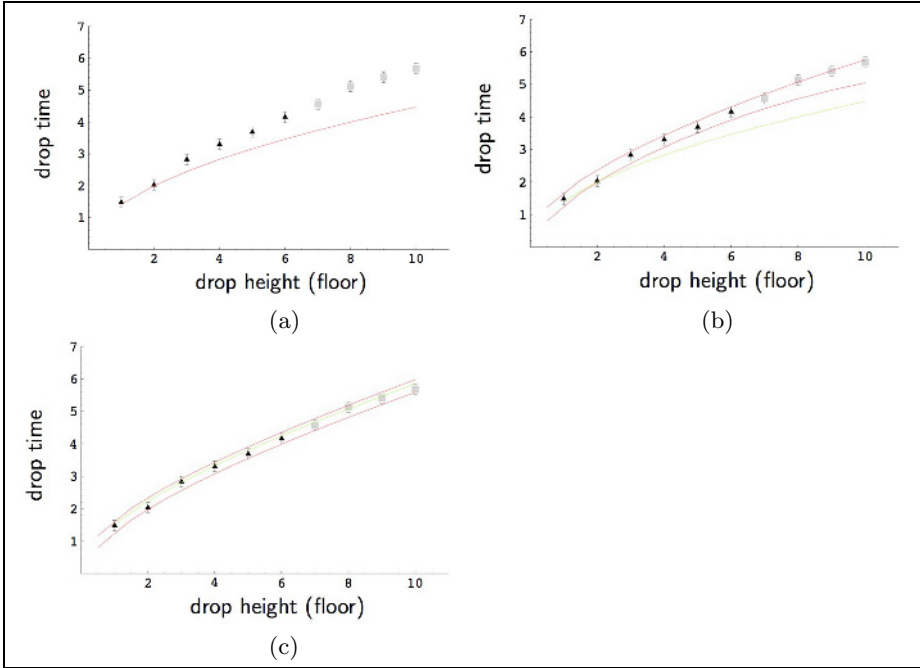


Fig. 1. Explanatory diagrams of experiment drop times. a) Results compared to an inadequate ideal model; b) model results and (bounded) discrepancy adjustment; c) revised model.

or simulation-based prediction, remains distinctly low, while the discrepancy adjustment, with uncertainty incorporated, follows the data closely and responds better in extrapolation.

Continuing the analysis, the discrepancy term suggests the postulation of an improved model, for example:

$$\frac{d^2z}{dt^2} = -1 - \theta \frac{dz}{dt} + \epsilon,$$

where θ is an unknown to be calibrated. Determining this model, including uncertainty both in the model and the calibration of the parameter, the results are shown in Fig. 1c. In this case, our best simulation result with the calibrated parameter closely follows the data, and the complete model closely bounds the prediction. The enhanced model gives greater confidence in extrapolation as compared to the incorrect model relying on estimated discrepancy to fit the data.

To summarize, the problem starts with the modeling of simulation response, and the modeling of experiments as the simulation response plus some systematic bias. This discrepancy describes simulation model insufficiency (or other systematic bias). Parameters of these models are determined so that uncertainty

in predictions is implicitly given. Unknown parameters are calibrated to best reconcile the experimental data with the simulations, reporting on plausible values. Discrepancy estimates may be further used to examine problem foundations.

2 Model Formulation

This modeling approach was originally generated by Kennedy and O'Hagan [6], and has been discussed in application in [8, 9]. For space limitations, the model cannot be completely described here, but complete formulation of the single experiment configuration methodology is available [10, 11].

The core of the modeling effort is a Gaussian stochastic process model (GPM). These models offer an expressive method for modeling a response over a generic space [12]. The GPM relies on a specified covariance structure, called the *covariogram*, that describes the correlation between data points that scales with distance. In this case, nearby is measured as a scaled Euclidean distance in the simulation parameter space. The covariogram is then:

$$C(x_1, x_2) = \frac{1}{\lambda_z} \exp^{-d(x_1, x_2, \beta)^2} + I \frac{1}{\lambda_s},$$

where $d = \sqrt{\sum \beta^i (x_1^i - x_2^i)^2}$. The λ parameters are precisions (inverse variances), with λ_z corresponding to the variability of the data, and λ_s corresponding to the variability of the residual. This correlation assumption constrains the response surface to a family functions. Predictions are made by computing a joint covariogram of the known and predicted datapoints, and producing the multivariate normal conditional distribution of the unknown locations. The predictions are then distributions rather than point estimates:

$$X_p \sim N(\mu_p, \Sigma_p).$$

This distribution can be used to produce realizations of possible values, and also allows the extraction of confidence bounds on the estimates.

η models the simulations, but a two-part model is used to also explicitly model the discrepancy δ between the observed data, y_{obs} , and the simulation response, y_{sim} , such that:

$$\begin{aligned} y_{sim} &= \eta(x, t) + \epsilon, \\ y_{obs} &= \eta(x, \theta) + \delta(x) + \epsilon. \end{aligned}$$

t are simulation parameters whose corresponding values are not known for the observed data. These unknown θ are determined (i.e., calibrated) in modeling, along with the β and λ parameters for for the η and δ models.

η and δ model parameters and θ values are generated with Markov chain Monte-Carlo sampling of a Bayesian posterior. The Bayesian approach gives the posterior density for the η model as

$$\pi(\cdot, \beta, \lambda | y_{sim}(x)) \propto L(y(x) | \eta(x, \beta, \lambda)) \times \pi(\beta) \times \pi(\lambda).$$

In words: the *posterior* distribution of the parameters given the simulation data is proportional to the likelihood of the data given the parameters times the prior probability of the parameters. The likelihood corresponds to a least squares measure of the model predictions compared to the given data, though this is not computed explicitly. This formulation has been extended to produce a single likelihood of the $\eta + \delta$ model fitting the observed and simulated data simultaneously [11].

The posterior of the parameters can be sampled from using Metropolis-Hastings MCMC sampler, resulting in samples from joint the posterior density of all parameters. It is possible to optimize directly on the likelihood function or the posterior for a point solution, but the resulting “optimal” solution has no associated confidence on the parameters. Computing the likelihood requires inversion of the $k \times k$ covariance matrix, where $k = n(p + q) + mq$, where n is the number of experimental data and m is the number of simulated data, and p is the dimensionality of the simulation response, and q is the dimensionality of the discrepancy response. This quickly becomes intractable if p grows large, so an important enhancement to the model is the use of linear basis dimension reduction in the output space, the effects of which can be compensated for in the modeling approach. Good results have been obtained with principle components reducing the p dimension, and kernel regression constraining and reducing q . This makes even problems with large data sizes, for example time series or images, computationally tractable [10].

3 Example Application: Flyerplate Experiments

In order to study the properties of materials under shock, plates of the material are subjected to a high-velocity impact. The velocity of the impacted plate is measured over time, revealing several material property driven regimes, as detailed in Fig. 2.

Results of flyerplate experiments are shown Fig. 3, which shows both the simulations from a 128 experiment design over 7 variables, and a trace of measured data. The unknown θ parameters ranges have been defined by subject matter experts, and scaled to $[0,1]$ for the purposes of the analysis. θ in this problem

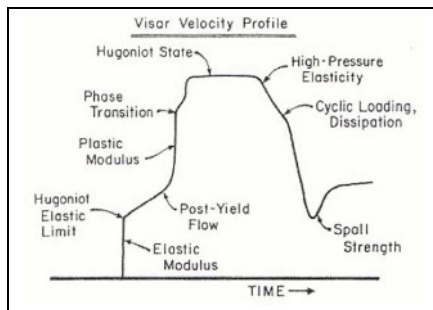


Fig. 2. Theoretical regions of flyerplate velocity measurements

are parameters from the Preston-Tonks-Wallace stress-strain model[5]. Parameters include: θ_0 Initial strain hardening rate; κ material constant in thermal activation energy; $-\log(\gamma)$ material constant in thermal activation energy; y_0 , maximum yield stress; y_∞ , minimum yield stress; s_0 , maximum saturation stress; s_∞ , minimum saturation stress. The simulation data is modeled with the first five principal components. The discrepancy model is a kernel basis constraining relatively smooth variation over the parameter space, modeling an arbitrary general effect in the absence of a more specific model.

Figure 4 shows the results of the full joint parameter calibration as contours of the two-dimensional PDF projections, as sampled by the MCMC procedure. There are clear trends in some variables, which have been calibrated more tightly than their *a priori* ranges, whereas some do not have clear preferred values. An

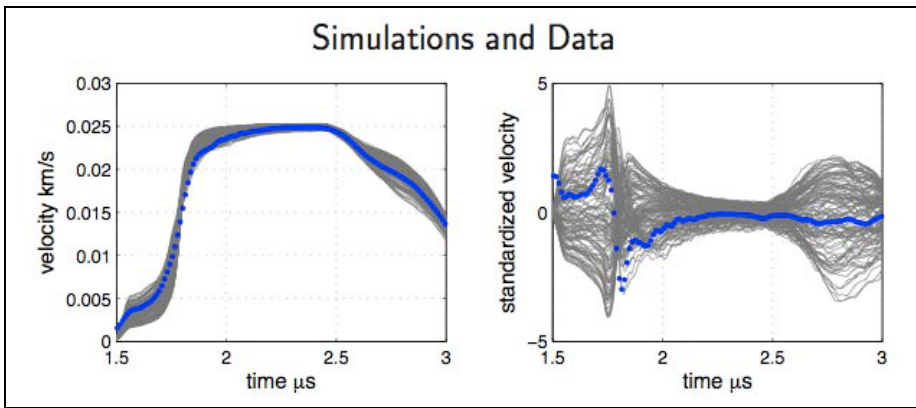


Fig. 3. Observed and simulated tantalum flyerplate velocity results, native data scale and standardized to mean 0 variance 1

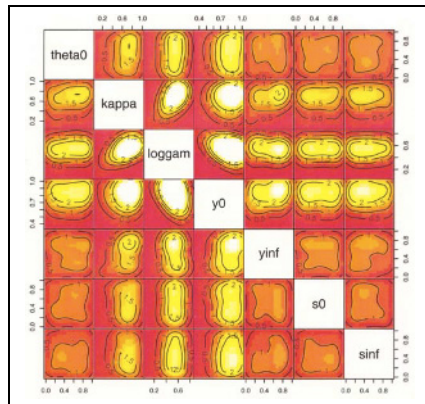


Fig. 4. Calibration results of unknown model parameters explored with simulations

additional variable importance measure comes from the sampled spatial scaling parameters β , where lower spatial correlation corresponds to more active variables. This experiment verified that β_2 , β_3 , and β_4 are important, consistent with the θ calibration. Proper sensitivity analysis can be performed by analyzing model predictions, which can be produced cheaply from the emulator model. Figure 5 shows predictive results of the model in the scaled space. These predic-

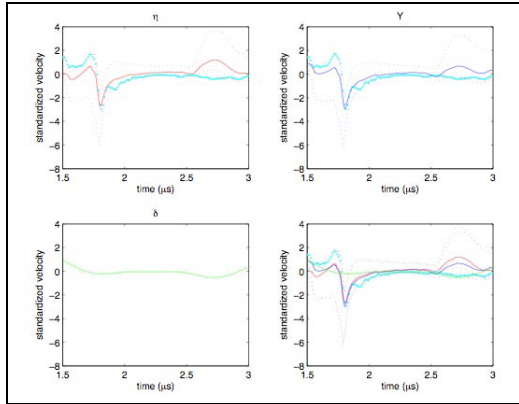


Fig. 5. Observed and calibrated predictions of the velocimetry results for tantalum flyerplate experiments (scaled). On the upper left is observed data in large dots, with the η model. Lower left shows the δ model contribution. The upper right is the observed data with the $Y = \eta + \delta$ result. The lower right plot repeats all results on one scale. Solid lines are mean response, and dotted lines show the associated 10%-90% confidence region of each quantity.

tions cover several model realizations (drawn from the Normal model distribution), predicted over many MCMC drawn parameter sets. The closest emulator simulation response is not able to capture the observed data, in particular in the edges where the simulations showed an initial higher-than-expected measurement, and late-time bump that is not observed in the data. The discrepancy adjusted response reduces this error. Of particular interest is that the simulator response alone not only failed to capture the observed data, but the confidence region is also not satisfactory. The $Y = \eta + \delta$ model's confidence regions are more appropriate.

4 Joint Calibration of Models

In complex applications, *separate effects test* are used to explore different aspects of a problem through surrogate experiments. Flyerplate results may be scientifically interesting in their domain, but the experiments are also intended to collect data on effects that are part of larger physical simulations. Some of these experiments will inform on the same parameters, and it is desired to perform a calibration that uses all of these results to simultaneously and consistently.

If two models have separate likelihoods $L_1(\theta_1, \theta_2|y)$ and $L_2(\theta_1, \theta_3|y)$, they can be considered a joint model as $L_J = L_1 \times L_2$. Using MCMC, the draws can be simplified, computing L_J for the draws related to θ_1 , while for the parameters of independent models the likelihoods, L_1 and L_2 are computed independently for draws of θ_2 and θ_3 , saving computation. This is a method to quantify the effects of variables in common between experiments.

In the *shock speed modeling* problem, it is desired to model measured shock speed quantities of several materials. A single experiment in a single material consists of shock particle velocity u_p measured in response to shock of speed u_s . Several experiments characterize a material response curve, which is modeled by simulations, as shown in Fig. 6.

In the full application, there are many materials and several parameters, some of which are shared between models. We will limit the discussion to the characterization of hydrogen (H) and deuterium (D), which use the same parameters, referred to here as θ_1 - θ_3 . Figure 7 shows the calibration of parameters for each single model, as well as the joint model. The joint model shows a more compact and stable calibration, in a region that is expected by domain experts. These

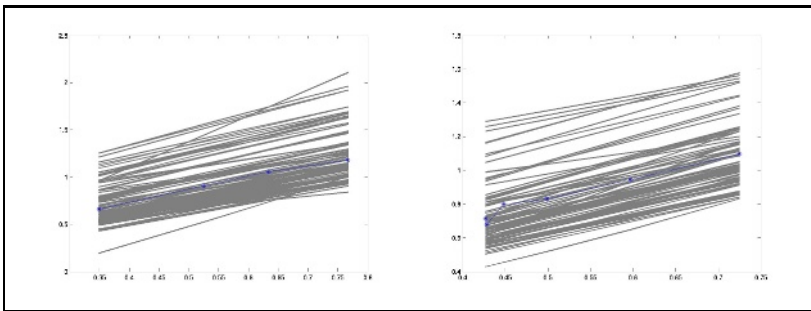


Fig. 6. Shock speed simulation models and measured data for hydrogen (left) and deuterium

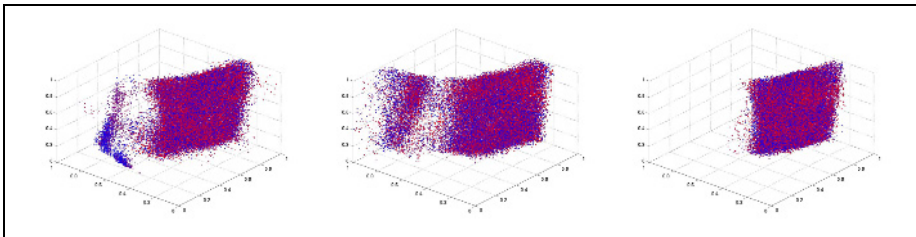


Fig. 7. Joint calibration of model parameters. The left plot shows the calibrated theta vectors for a hydrogen model alone, the middle plot shows the same parameters in a deuterium model, and the right plot shows the joint calibration of the parameters given both datasets.

results show that this methodology can successfully capture data from different experiments and even different simulation systems to calibrate underlying parameters with greater fidelity than the single model calibrations are capable.

5 Discussion

The approach described provides a method to quantify the implications of experimental data combined with simulation experiments. It is a tool to be used with domain experts from both the modeling and simulation domain, as well as the experimental data domain. Expert judgement is required in the generation of appropriate simulation studies, the construction of plausible discrepancy models, and the assessment of results. If domain knowledge is available to describe strong parameter priors, including θ parameter bounds and relationships, these may be incorporated, though by default weak priors that do not inappropriately constrain modeling and calibration can be used.

As is usual in complex models, attention to diagnostics is important. Because this modeling approach incorporates a number of trade-offs in the modeling space, it is possible that the model could enter an inappropriate domain, where it is fitting the data well, but the parameters are not plausible (e.g., counter to physics). Also, some expertise in MCMC is needed to identify the initial transient and to determine appropriate step sizes over the parameter ranges to ensure adequate mixing of the chain.

Without response smoothness, it is difficult to envision how to model and calibrate in this (or any) framework. Thus, a key issue in successful application is ensuring simulation response smoothness through the parameter space under analysis. If this is not an inherent quality of the data, variable transformations and feature construction studies are necessary.

In summary, this modeling approach:

- provides a method for integrating simulation experiments with empirical data, modeling systematic error in the simulation;
- calibrates unknown simulation parameters;
- provides well-founded uncertainty estimates in parameters and predictions; and
- allows separate experiment results to be fused into one result of parameter calibration.

When given two distinct datasets with a relationship in their underlying variables, it is generally not clear how to fuse the information in the datasets into a single answer. Information must be of the same type before it can be quantitatively reconciled, and this is usually solved by transforming the data directly into the same domain. The application methodology described here shows how different datasets may be linked by a generating model, in this case a simulation that can produce results in the various model domains. Through this approach, inverse problems from two distinct experimental domains can be combined, and a composite model realized.

References

1. *Los Alamos Science* special issue on Science-Based Prediction for Complex Systems, no.29, 2005.
2. J. Gattiker, "Using the Gaussian Process Model for Simulation Analysis Code", Los Alamos technical report LA-UR-05-5215, 2005.
3. Christie, Glimm, Grove, Higdon, Sharp, Schultz, "Error Analysis in Simulation of Complex Phenomena", *Los Alamos Science* special issue on Science-Based Prediction for Complex Systems, no.29, 2005, pp.6-25.
4. S. Chib, E. Greenberg, "Understanding the Metropolis-Hastings Algorithm", *The American Statistician*, Nov. 1995; 49, 4; p.327.
5. M Fugate, B Williams, D Higdon, K Hanson, J Gattiker, S Chen, C Unal, "Hierarchical Bayesian Analysis and the Preston-Tonks-Wallace Model", Los Alamos Technical Report, LA-UR-05-3935, 2005.
6. M Kennedy, A. O'Hagan, "Bayesian Calibration of Computer Models (with discussion)", *Journal of the Royal Statistical Society B*, 68:426-464.
7. D.Jones, M. Schonlau, W.Welch, "Efficient Global Optimization of Expensive Black-Box Functions", *Journal of Global Optimization* 13, pp. 455-492, 1998.
8. D. Higdon, M. Kennedy, J. Cavendish, J. Cafoe, and R. Ryne, "Combining field Observations and Simulations for Calibration and Prediction", *SIAM Journal of Scientific Computing*, 26:448-466.
9. C. Nakhleh, D. Higdon, C. Allen, V. Kumar, "Bayesian Reconstruction of Particle Beam Phase Space from Low Dimensional Data", Los Alamos technical report LA-UR-05-5897.
10. Dave Higdon, Jim Gattiker, Brian Williams, Maria Rightley, "Computer Model Calibration using High Dimensional Output", Los Alamos Technical report LA-UR-05-6410, submitted to the *Journal of the American Statistical Association*.
11. Brian Williams, Dave Higdon, James Gattiker, "Uncertainty Quantification for Combining Experimental Data and Computer Simulations", Los Alamos Technical Report LA-UR-05-7812.
12. D.R.Jones, M.Schonlau, W.Welch, "Efficient Global Optimization of Expensive Black-Box Functions", *Journal of Global Optimization* 13, pp. 455-492, 1998.

A Fast Method for Detecting Moving Vehicles Using Plane Constraint of Geometric Invariance*

Dong-Joong Kang¹, Jong-Eun Ha², and Tae-Jung Lho¹

¹ Dept. of Mechatronics Eng., Tongmyong University,
535, Yongdang-dong, Nam-gu, Busan 608-711, Korea
{dj kang, tj lho}@tit.ac.kr

² Dept. of Automotive Eng., Seoul National University of Technology,
138, Gongrung-gil, Nowon-gu, Seoul 139-743, Korea
jeha@snut.ac.kr

Abstract. This paper presents a new method of detecting on-road highway vehicles for active safety vehicle system.* We combine a projective invariant technique with motion information to detect overtaking road vehicles. The vehicles are assumed into *a set of planes* and the invariant technique extracts the plane from the theory that a geometric invariant value defined by five points on a plane is preserved under a projective transform. Harris corners as a salient image point are used to give motion information with the normalized cross correlation centered at these points. A probabilistic criterion without demand of a heuristic factor is defined to test the similarity of invariant values between sequential frames. Because the method is very fast, real-time processing is possible for vehicle detection. Experimental results using images of real road scenes are presented.

Keywords: Geometric invariant, plane constraint, motion, vehicle detection, active safety vehicle.

1 Introduction

There are growing social and technical interests in developing vision-based intelligent vehicle systems for improving traffic safety and efficiency. Intelligent on-road vehicles, guided by computer vision systems, are a main issue in developing experimental or commercial vehicles in numerous places in the world [1-7].

Reliable vehicle detection in images acquired by a moving vehicle is an important problem for the applications such as active safety vehicles (ASV) equipped with driver assistance system to avoid collision and dangerous accidents. Several factors including changing environmental condition affect on-road vehicle detection and the appearance changes of foregoing vehicles with scale, location, orientation, and pose transition makes the problem very challenging.

Foregoing vehicles are come into several views with different speeds and may always vary in shape, size, and color. Vehicle appearance depends on relative pose

* This work was supported by Tongmyong Univ. of Information Tech. Research Fund of 2005.

between observer and foregoing vehicles and occlusion by nearby objects affects the detection performance. In case of real implementation of intelligent road vehicle, real-time processing is another important issue.

We consider the problem of rear-view detection of foregoing and overtaking vehicles from gray-scale images. Several previous researches assume two main steps to detect road vehicles [1]. The first step of any vehicle detection system is hypothesizing the locations in images where vehicles are present. Then, verification is applied to test the hypotheses. Both steps are equally important and challenging. Well known approaches to generate the locations of vehicles in images include using motion information, symmetry, shadows, and vertical/horizontal edges [4-7].

The purpose of this paper is to provide a method for the hypothetical candidates of on-road vehicle detection. Once the hypothetical regions including vehicles are extracted first, then several methods could be applied to verify the initial detection.

Detecting moving objects from images acquired by a static camera can be usually performed by simple image difference based methods. However, when the camera undergoes an arbitrary motion through a scene, the task is much more difficult since the scene is no longer static in the image. Simple image differencing techniques no longer apply. Road vehicle detection problem belongs to the second category because the observer camera is mounted on a moving vehicle. For general segmentation, optical flow from all image points or corresponding information from prominent image features can be used.

In this paper, we present a method based on the projective invariant and motion information. Based on the sparsely obtained motion field or corresponding data, the method selects initial segmentation clusters by using projective invariant method that can be described by a plane constraint. Moving vehicle segmentation is based on the fact that a geometric invariant value of point-set defined on a plane of the vehicle is preserved after motion of the plane [8-10]. Harris corner [11] is a good image feature to provide motion information with the normalized correlation. The probabilistic criterions to test the similarity of invariant values between frames are introduced without a need of magic factors for the threshold. The proposed method is more exact in initial segmentation than simple methods clustering similar motion vectors because a side or rear part of a vehicle could be separately extracted under the *strong plane constraint*. The method is very fast and the processing time of each module is evaluated.

Among the vehicles surrounding host vehicle, close-by front and rear, and overtaking side vehicles are more dangerous for collision and threaten car driver [1]. Methods detecting vehicles in these regions might be better to employ motion information because there are large intensity changes and detailed image features such as edges and corners by close view. Vehicles in the far distance region are relatively less dangerous and appearance is more stable since the full view of a vehicle is available.

2 Segmentation of Planes on Moving Vehicle

2.1 Projective Invariants

Projective invariants are quantities which do not change under projective transformations. Detailed contents of the uses of invariants are given in Mundy and

Zisserman [8]. There are two convenient invariants that can be defined for groups of five points. Four points (no three collinear) form a projective basis for the plane and the invariants correspond to the two degrees of freedom of the projective position of the fifth point with respect to the first four - there exist positions that invariants do not change their values in some directions. Fig. 1 shows the five-point sets on a plane under arbitrary projective transformation. The invariant value defined by five points on the plane is not changed under a projective transformation.

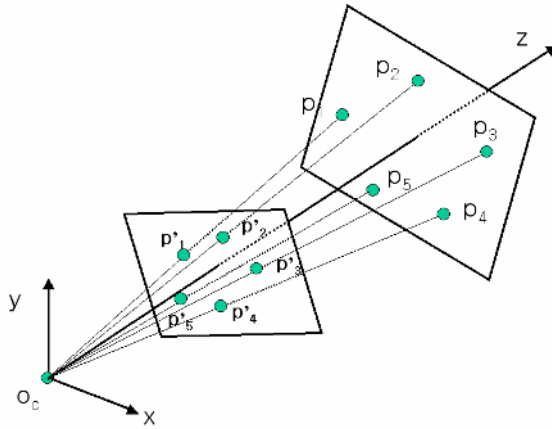


Fig. 1. Five point invariance on a plane

The two invariants may conveniently be written as the ratios of determinants of matrices of the form \mathbf{M}_{ijk} , which denotes area of a triangle consisting of three image points. Then the two invariants are given by [8-9]:

$$\mathbf{I}_1 = \frac{|\mathbf{M}_{124}||\mathbf{M}_{135}|}{|\mathbf{M}_{134}||\mathbf{M}_{125}|} \tag{1}$$

$$\mathbf{I}_2 = \frac{|\mathbf{M}_{241}||\mathbf{M}_{235}|}{|\mathbf{M}_{234}||\mathbf{M}_{215}|} \tag{2}$$

where $\mathbf{M}_{ijk} = (\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$ and \mathbf{x}_i is position (x_i, y_i) of an image point. These two quantities may be seen to be preserved under a projective transformation if \mathbf{x}' is substituted for \mathbf{x} ,

$$\frac{|\mathbf{M}'_{124}||\mathbf{M}'_{135}|}{|\mathbf{M}'_{134}||\mathbf{M}'_{125}|} = \frac{|\lambda_1 \mathbf{P}\mathbf{x}_1, \lambda_2 \mathbf{P}\mathbf{x}_2, \lambda_4 \mathbf{P}\mathbf{x}_4||\lambda_1 \mathbf{P}\mathbf{x}_1, \lambda_3 \mathbf{P}\mathbf{x}_3, \lambda_5 \mathbf{P}\mathbf{x}_5|}{|\lambda_1 \mathbf{P}\mathbf{x}_1, \lambda_3 \mathbf{P}\mathbf{x}_3, \lambda_4 \mathbf{P}\mathbf{x}_4||\lambda_1 \mathbf{P}\mathbf{x}_1, \lambda_2 \mathbf{P}\mathbf{x}_2, \lambda_5 \mathbf{P}\mathbf{x}_5|} \tag{3}$$

which gives,

$$\frac{|\mathbf{M}'_{124}||\mathbf{M}'_{135}|}{|\mathbf{M}'_{134}||\mathbf{M}'_{125}|} = \frac{\lambda_1^2 \lambda_2 \lambda_3 \lambda_4 \lambda_5 |\mathbf{P}|^2 |\mathbf{M}_{124}||\mathbf{M}_{135}|}{\lambda_1^2 \lambda_2 \lambda_3 \lambda_4 \lambda_5 |\mathbf{P}|^2 |\mathbf{M}_{134}||\mathbf{M}_{125}|} \tag{4}$$

where \mathbf{P} is the projectivity matrix and λ_i is scaling factor.

2.2 Extract Point Groups from Motion Data

Point features corresponding to high curvature points are extracted from image before motion [11-13]. A salient image feature should be consistently extracted for different views of object and there should be enough information in the neighborhood of the feature points so that corresponding points can be automatically matched. A popular method for corner detection is Harris detector [11] using convolution operation, in which the method obtains the high curvature matrices related to the convolution for partial derivatives of image data.

In notations of Harris corner detector, the g in eq. (6) denotes gray scale image and w is the Gaussian smoothing operator, k is an experimental parameter. And g_x and g_y indicates the x and y directional derivative for the grey image, respectively. Corners are defined as local maxima of the corner response function \mathbf{R} :

$$\mathbf{R}(x, y) = \det[\mathbf{C}] - k \cdot \text{trace}^2[\mathbf{C}] \tag{5}$$

where \mathbf{C} is

$$\mathbf{C} = w \cdot \begin{bmatrix} g_x^2 & g_x g_y \\ g_x g_y & g_y^2 \end{bmatrix}. \tag{6}$$

Given the high curvature points, we can use a correlation window of small size $(2r + 1) \times (2c + 1)$ centered at these points. A rectangle search area of size $(2d_x + 1) \times (2d_y + 1)$ is defined around the points in the next image, and the intensity-normalized cross correlation (NCC) [14] on a given window is performed between a point in the first image and pixels within the search area in the second image. This operation provides the matched motion vectors.

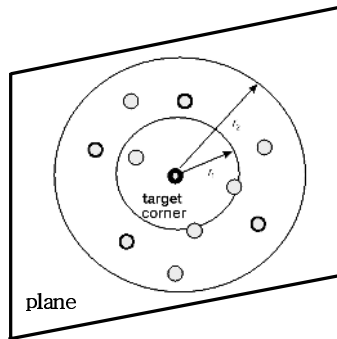


Fig. 2. Selection of four points defining invariant values around a center point

Selection of five points from n motion data is needed to define invariant values. There are $N = {}_n C_5$ independent ways of choosing five points from n . Therefore, it is impractical to test all possible combinations of five points in an image.

Instead of, groups of five points are selected as four nearest neighbors outside a small circular neighborhood of the fifth point and inside of a larger circle, as shown in Fig. 2. This gives only small groups of points to be tested and reduction of processing time is possible. In the circular band region, four points are selected from random combination of the points in the band. Among the combination, the candidates that give three collinear points or have the points that close to each other are rejected.

2.3 A Probabilistic Criterion to Decide Point Groups on a Plane

Five-point sets are randomly selected in the circular band to define projective invariant values. A set before motion defines a model invariant value and the corresponding set after motion defines the corresponding invariant value. If this pair exists on the same plane or on an object moving under weakly perspective projection assuming far distance from the observer, the difference of two invariant values will be small. This constraint makes the plane segmentation possible. Pairs of points giving a similar invariant value between after and before motion are considered as side planes of independently moving objects. We introduce a threshold to test the similarity of two invariant values.

$$|\mathbf{I}'_i - \mathbf{I}_i| < Thres_i \tag{7}$$

For selecting thresholds, a probabilistic criterion is introduced by a measurement uncertainty associated with the estimated position of corner features [9-10].

The invariant is a function of five points:

$$\mathbf{I} = \mathbf{I}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5). \tag{8}$$

Let \mathbf{x}_i be the true and $\tilde{\mathbf{x}}_i$ be the noisy observation of (x_i, y_i) , then we have

$$\tilde{x}_i = x_i + \xi_i \tag{9a}$$

$$\tilde{y}_i = y_i + \eta_i \tag{9b}$$

where the noise terms ξ_i and η_i denote independently distributed noise terms having mean 0 and variance σ_i^2 . From these noisy measurements, we define the noisy invariant,

$$\tilde{\mathbf{I}}(\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \tilde{\mathbf{x}}_3, \tilde{\mathbf{x}}_4, \tilde{\mathbf{x}}_5) \tag{10}$$

To determine the expected value and variance of $\tilde{\mathbf{I}}$, we expand $\tilde{\mathbf{I}}$ as a Taylor series at $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5)$:

$$\mathbf{I} \approx \tilde{\mathbf{I}} + \sum_{i=1}^5 \left[(\tilde{x}_i - x_i) \frac{\partial \tilde{\mathbf{I}}}{\partial \tilde{x}_i} + (\tilde{y}_i - y_i) \frac{\partial \tilde{\mathbf{I}}}{\partial \tilde{y}_i} \right] = \tilde{\mathbf{I}} + \sum_{i=1}^5 \left[\xi_i \frac{\partial \tilde{\mathbf{I}}}{\partial \tilde{x}_i} + \eta_i \frac{\partial \tilde{\mathbf{I}}}{\partial \tilde{y}_i} \right] \tag{11}$$

Then, the variance becomes

$$E[(\tilde{\mathbf{I}} - \mathbf{I})^2] = \sigma_0^2 \sum_{i=1}^5 \left[\left(\frac{\partial \tilde{\mathbf{I}}}{\partial \tilde{x}_i} \right)^2 + \left(\frac{\partial \tilde{\mathbf{I}}}{\partial \tilde{y}_i} \right)^2 \right]. \tag{12}$$

Hence, for a given invariant \mathbf{I} , we can determine a threshold:

$$\Delta \mathbf{I} = 3 \cdot \sqrt{E[(\tilde{\mathbf{I}} - \mathbf{I})^2]}. \tag{13}$$

The partial derivative $\partial \mathbf{I}_1 / \partial x_1$, for example, is given by

$$\frac{\partial \mathbf{I}_1}{\partial x_1} = \mathbf{I}_1 \cdot \left[\frac{y_2 - y_4}{|\mathbf{M}_{124}|} + \frac{y_3 - y_5}{|\mathbf{M}_{135}|} + \frac{y_3 - y_4}{|\mathbf{M}_{134}|} + \frac{y_2 - y_5}{|\mathbf{M}_{125}|} \right]. \tag{14}$$

Because there are two different invariant values, we have to define two threshold values $\Delta \mathbf{I}_1$ and $\Delta \mathbf{I}_2$. If there are point groups smaller than two threshold values defined in the circular band region, the local region by the point set is selected as plane candidate on a moving vehicle.

3 Experiments

Experiments show overtaking close-by vehicles are well detected. Fig. 3 presents a detection example for real road scene. Corners on the rear and side part of a vehicle are extracted by Harris corner algorithm as shown in Fig. 3(a). We set $r_1 = 3$ and $r_2 = 15$ pixels to define the circular band of Fig. 2 for calculation of the geometric invariant values. Image size and processing region is 256x256 and 232x150 pixel², respectively.

Fig. 3(b) shows motion vectors from the normalized cross correlation employed to Harris corners of Fig. 2(a) between two sequential frames. The motion vectors in Fig. 3(b) are magnified to 3 times for good visualization. The correlation is performed with small size window of 7x7 pixel² ($r = c = 3$) for search regions of 11x11 pixel² ($d_x = d_y = 5$) for the image of next frame. Points giving a small difference of invariant value between two frames are showed as the small white boxes in Fig. 3(c), and Fig. 3(d) shows each MBR from five point sets. Nine regions on side planes are found after no motion MBRs are rejected if exists. Position variance σ_o in eq. (12) sets to 0.2. As the center points of detected five point sets are appeared on side of vehicle, the side of a vehicle is recognized as an approximate planar object with the invariant values preserved during moving of the vehicle. The range of the automatic threshold value $\Delta \mathbf{I}_1$ obtained during processing is between about 0.06~2.0 in case of Fig. 4(d).

Fig. 4 shows rear and side parts of a vehicle are separated as two different planes for sequential image frames. The two planes are consistently detected during 8 frames

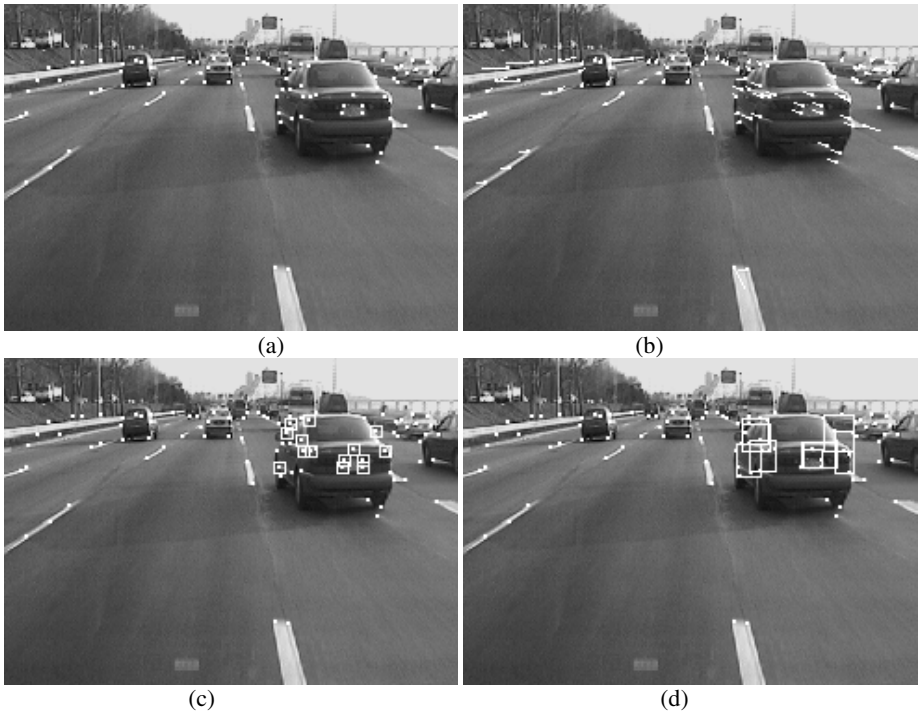


Fig. 3. Extraction of side planes on a overtaking vehicle. (a) Harris corners; (b) Motion vectors by NCC; (c) Center corners included in invariant planes; (d) Detection of vehicle planes.

of sequential images. Small white rectangles in Fig. 4 present detection of rear and side planes of a moving vehicle. The regions with smaller motion than 3 pixels in average of five points in the white region are rejected to prevent noisy extraction of far distance vehicles and background regions.

Different parameter of Harris detector as change of the threshold value presents different level of corner detection. Table 1 shows the resulting computing time for two different corner detection images. The processing image region is $232 \times 150 (=w \times h)$ pixels² and we use Pentium-IV 2.0Ghz processor under Windows-XP environment. We do not use any optimization procedures such as Intel SIMD, MMX technologies. The total elapsed time for two cases is 39.1 and 40.1 msec, respectively.

Table 1. Computing time for each module of the proposed algorithm

	Case A (corner #: 202)	Case B (corner #: 255)
SUSAN detector	31 msec	31 msec
Motion calculation	5.1 msec	6.3 msec
Invariant plane detection	3 msec	3.6 msec
Total elapsed time	39.1 msec	40.9 msec

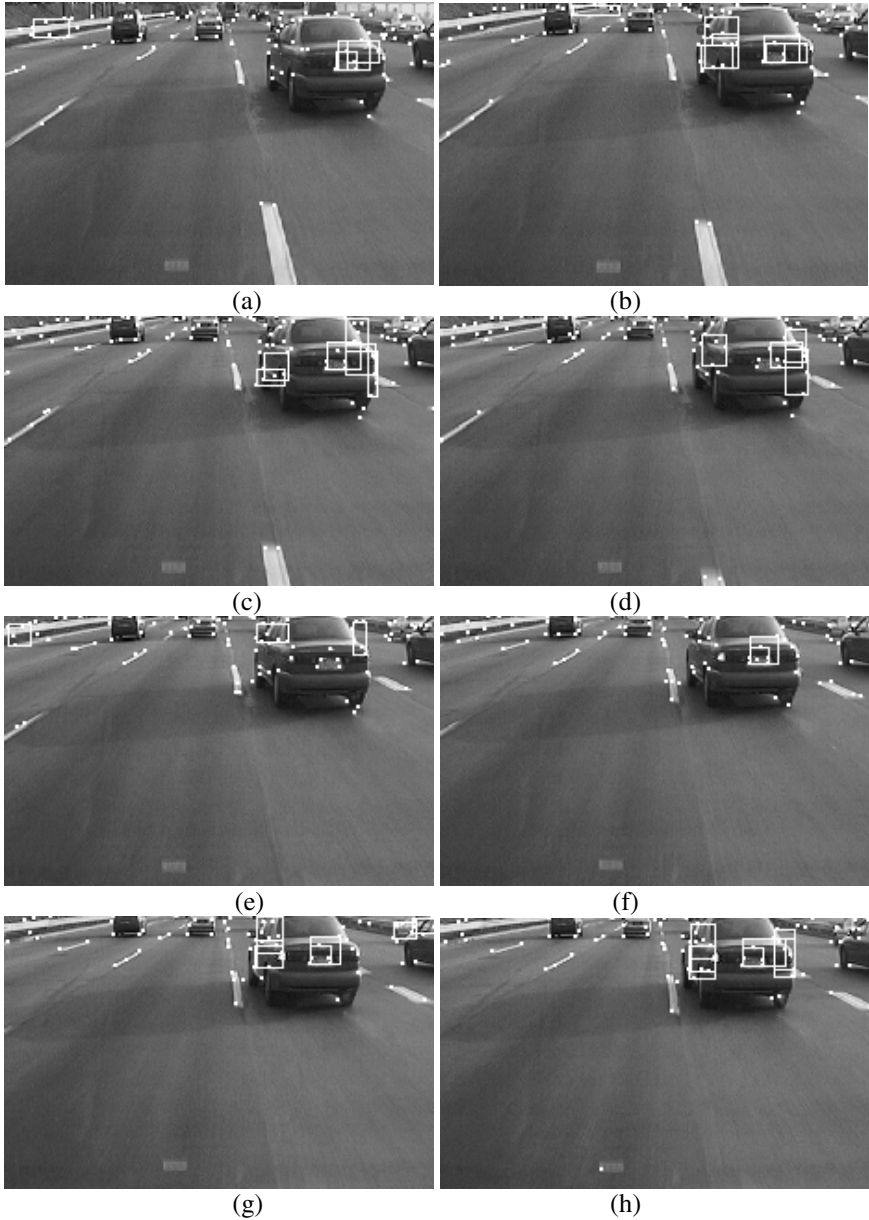


Fig. 4. Side plane detection of close-by overtaking vehicle on highway for 8 sequential frames during about 0.2 sec

4 Conclusions

We present a geometric invariant and motion based method for vision-based vehicle detection. The method uses the projective invariant value from the motion information

obtained on corner points of sequential frames. The proposed method assumes the 3-D moving vehicle into a set of planar surfaces that come from rear and side surface parts of a vehicle. The method is more exact in the plane segmentation than the clustering methods using similarity of motion vectors because side or rear parts of a vehicle could be independently extracted under strong plane constraint of projective invariance. The five point invariant values can prove detection of real planes on a targeted vehicle.

For consecutive images, a corner extraction algorithm detects prominent corner features in the first image. Correspondence information of detected corner points is then found by the normalized correlation method for next sequential image. Based on the sparsely obtained motion data, the segmentation algorithm selects points on a plane that maintain consistent projective invariants between frames. These points set can form initial clusters to segment the plane and other post-procedures might be applied to merge the neighboring points having a similar motion. Each processing module is very fast and adequate to real-time processing for fast moving highway vehicles even though many processing units including corner detector, NCC matching, random sampling-based five point selector, and probabilistic invariant plane extraction are introduced. Through the experiments of real road scenes, the proposed method presents a plane-by-plane segmentation of fast moving and overtaking vehicles is possible.

References

1. Sun, Z., Bebis G., Miller R.: On-road vehicle detection using optical sensors: A review, IEEE Intelligent Transportation Systems Conference, Washington D.C., (2004) 585-590
2. A. Polk and R. Jain, I.Masaki: Vision-based Vehicle Guidance, Springer-Verlag, (1992)
3. Kang, D.J., Jeong M.H.: Road lane segmentation using dynamic programming for active safety vehicles, Pattern Recognition Letters, Vol.24 (2003) 3177-3185
4. Smith, S.M.: ASSET-2 Real-Time Motion Segmentation and Object Tracking, Int. Conf. on Computer Vision, (1995) 237-244
5. Kuehnle, A.: Symmetry-based recognition for vehicle rears, Pattern Recognition Letters, Vol.12 (1991) 249-258
6. Handmann, U., Kalinke, T., Tzomakas, C., Werner, M., Seelen, W.: An image processing system for driver assistance, Image and Vision Computing, Vol. 18 (2000) 367-376
7. Betke, M., Haritaglu, E., Davis, L.: Multiple vehicle detection and tracking in hard real time, IEEE Intelligent Vehicles Symposium, (1996) 351-356
8. Joseph, L., Mundy, Zisserman, A.: Geometric Invariance in Computer Vision, MIT Press, (1992)
9. Roh, K.S., Kweon, I.S.: 2-D Object Recognition Using Invariant Contour Descriptor and Projective Refinement, Vol. 31, (1988) 441-455
10. Sinclair, D., Blake A.: Quantitative Planar Region Detection, Int. Journal of Computer Vision, Vol.18, (1996) 77-91
11. Harris, C., Stephens, M.: A combined corner and edge detector, The Fourth Alvey Vision Conference, (1988) 147-151
12. Press, W., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P.: Numerical Recipes in C, Cambridge University Express, (1992)
13. Smith, S.M., and Brady, J.M.: SUSAN - a new approach to low level image processing, Int. Journal of Computer Vision, Vol.23(1) (1997) 45-78
14. Gang, X., Zhang, Z.: Epipolar geometry in Stereo, Motion and Object Recognition, Kluwer Academic Publishers, (1996) 223-237

Robust Fault Matched Optical Flow Detection Using 2D Histogram

Jaechoon Chon¹ and Hyongsuk Kim²

¹ Center for Spatial Information Science at the University of Tokyo,
Cw-503 4-6-1 Komaba, Meguro-ku, Tokyo
jcchon@iis.u-tokyo.ac.jp

² Chonbuk National University, Korea
hskim@chonbuk.ac.kr

Abstract. This paper propose an algorithm by which to achieve robust outlier detection without fitting camera models. This algorithm is applicable for cases in which the outlier rate is over 85%. If the outlier rate of optical flows is over 45%, then discarding outliers with conventional algorithms in real-time applications is very difficult. The proposed algorithm overcomes conventional difficulties by using a three-step algorithm: 1) construct a two-dimensional histogram with two axes having the lengths and directions of the optical flows; 2) sort the number of optical flows in each bin of the two-dimensional histogram in descending order, and remove bins having a lower number of optical flows than the given threshold; 3) increase the resolution of the two-dimensional histogram if the number of optical flows grouped in a specific bin is over 20%, and decrease the resolution if the number of optical flows is less than 10%. This process is repeated until the number of optical flows falls into a range of 10%-20%. The proposed algorithm works well on different kinds of images having many outliers. Experimental results are reported.

1 Introduction

Using an image sequence to generate graphical information, such as image mosaics, stereoscopic imagery, 3D data, and moving object tracking, requires the use of optical flows. Presently, active studies on optical flow detection algorithms can be categorized into three different approaches: gradient-based [1], frequency-based [2], and feature-based [3]. In the gradient-based approach, the optical flow of each pixel is computed using a spatial and temporal gradient under the assumption of constant intensity in local areas [4]. In the frequency-based approach, the velocity of each pixel is computed by detecting the trajectory slope with velocity-tuned band-pass filters, such as Gabor filters [2].

The optical flows obtained from the plain area of an image are often incorrect and are not suitable for practical or real-time applications. Fortunately, to produce image mosaicking, epipolar geometry, and camera motion estimations, determining the optical flow at every point is unnecessary; in fact, ascertaining a small number of correct optical flows is more useful, and the feature-based approach is effective for making such determinations. Generally, feature points are extracted by algorithms that detect

corner points. Current corner-point-detection algorithms are referred to as Plessey [5], Kitchen/Rosenfeld [6], Curvature Scale Space [7], and Smallest Univalued Segment Assimilating Nucleus (SUSAN) [8]. The SUSAN algorithm used in this paper is an efficient feature-point-detection algorithm, utilizing the strong Gaussian operation of the intensity difference in the spatial domain. The feature-based approach requires feature point matching. In the common approach to match feature points as a pair, the first step is to take a small region of pixels surrounding the feature point to be matched between two frames. Then, compare the matched feature point with a similar window around each of the potential matching feature points in the other image. Each comparison yields a score related to the measure of similarity between the feature points by examination of the highest matching rate representing the best match. The measure of similarity can be evaluated via several ways, such as Normalized Cross-Correlation (*NCC*), Sum of Squared Difference (*SSD*), and Sum of Absolute Difference (*SAD*). The *NCC* used in this paper is the standard statistical method to determine similarity.

After a process of feature point matching or optical flow detection, robust regression algorithms are needed to discard incorrect optical flows; such incorrect optical flows result from many sources, such as image blur and changes caused by camera vibration, stereo motion, or zoom motions. Let correct and incorrect optical flows be inlier and outlier, respectively. The representative algorithms of robustly discarding outliers are E-estimators [9], Least Median of Square (LMedS) [9],[10], Random Sample Consensus (RANSAC) [11], and Tensor Voting [12]. The RANSAC algorithm is generally used to discard the outliers and to select the eight best-detected optical flows for robust fitting epipolar geometry [13]. In contrast to RANSAC, Tensor Voting can discard outliers without fitting camera models. The input of optical flows is first transformed into a sparse 8D point set. Dense, 8D tensor kernels are then used to vote for the most salient hyper plane that captures all inliers inherent in the input. With this filtered optical flows, the normalized eight-point algorithm can be used to estimate the fundamental matrix accurately. Meanwhile, probability of choosing inliers with well fitting models related with camera motion is so higher than that by E-estimator and LMedS, Tensor Voting and RANSAC are away from real-time application. In contrast to Tensor Voting and RANSAC, the E-estimator and LMedS can be applied to real-time computation. If the outlier rate among the detected optical flows is over 45%, discarding the outliers with LMedS is very difficult. For the E-estimator, it is impossible to apply optical flows that include over 35% outliers.

To overcome these limitations, we propose a robust outlier detection algorithm without fitting camera models for real-time applications and for cases in which the outlier rate is over 85%. When detecting optical flows between two input images, inliers in local areas have similarities, such as the similar directions and lengths of optical flows. In contrast, outliers have no similarities. Even if there are keeping the similarity, the number of outliers with the similarity is very few. Using basic concept of the different point, we propose an algorithm for grouping similar optical flows in a two-dimensional histogram that has two axes for the lengths and directions of optical flows and for removing some bins with lower numbers of optical flows. The proposed algorithm of detecting inliers in the two-dimensional histogram comprises the following three-steps: 1) construct a two-dimensional histogram having two axes for the lengths and directions of optical flows; 2) sort the number of optical flows in each bin

of the two-dimensional histogram by descending order, and remove some bins with lower numbers of optical flows than the threshold; 3) increase the resolution of the two-dimensional histogram if the number of optical flows grouped in a specific bin is over 20%, and decrease the resolution if the number of optical flows is less than 10%. This process is repeated until the number of optical flows falls into the range of 10%-20%. After completing the proposed three-steps, we can directly apply the chosen inliers to a computation of epipolar geometry by RANSAC to save computation costs, image mosaicking, moving objects tracking under a moving camera, and so on.

2 Outlier Detection Using Two-Dimensional Histogram

Optical flows, such as the vectors shown in Fig. 1(a), may be detected via processes of feature points extraction and matching without supporting any image motion information. The detected optical flows are composed of direction, θ , and length, l , as shown in Fig. 1(b). When sorting the number of optical flows into a two-dimensional histogram with two axes for the directions and lengths of the optical flows, the result can be depicted as shown in Fig. 1(c).

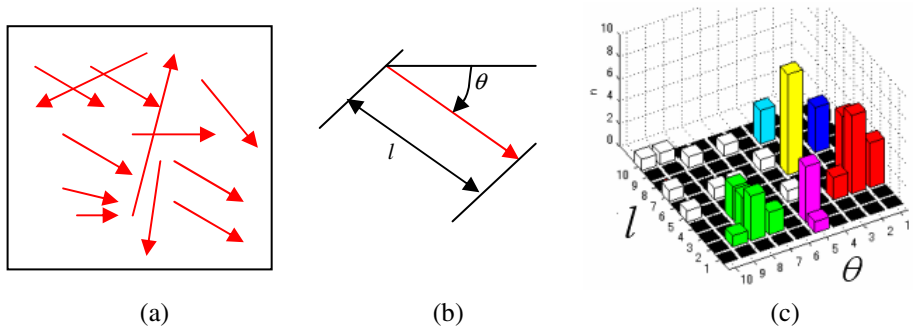


Fig. 1. Optical flows and their two-dimensional histogram. (a) optical flows, (b) the magnitude and directional elements of an optical flow, (c) two-dimensional histogram

The proposed algorithm for detecting inliers on the created two-dimensional histogram is composed of three steps shown in the flow chart in Fig. 2.

Step 1 is to construct a two-dimensional histogram with two axes for the lengths and directions of the optical flows. Step 2 is to sort the number of optical flows grouped in each bin of the two-dimensional histogram in descending order and to remove some bins having lower numbers of optical flows than the threshold. Step 3 is to increase the resolution of the two-dimensional histogram if the number of optical flows in a specific bin is over 20% and to decrease the resolution if the number of optical flows is less than 10%. This process is repeated until the number of optical flows falls into the range of 10%-20%.

When the percentage of optical flows that are inliers over is 40%, we can obtain good results, even if setting the upper limit of the range as a higher value. If the percentage is lower than 30%, then good results are not guaranteed. Therefore, to successfully apply the proposed algorithm to any kind of input images, the upper limit is

set as 20%. If the lower limit of the range is set as a so lower value and optical flows are detected on images taken from a camera with optical axis motion, then most optical flows will be discarded as outliers, because the number of locally grouped optical flows is too low. Consequently, to detect a stable number of optical flows from these images, the lower limit is set as 10%.

If the total number of optical flows grouped in all bins is over 20% of all initial optical flows and, simultaneously, that grouped in a specific bin with the highest number of optical flows is over 35% of the optical flows grouped in all bins, then the input images are taken only from a camera with translational motion, except for zoom motions. Of course, when the distance between objects and a panning and titling camera is far, the phenomenon will be generated. In these cases, only optical flows grouped in the specific bin are chosen as inliers.

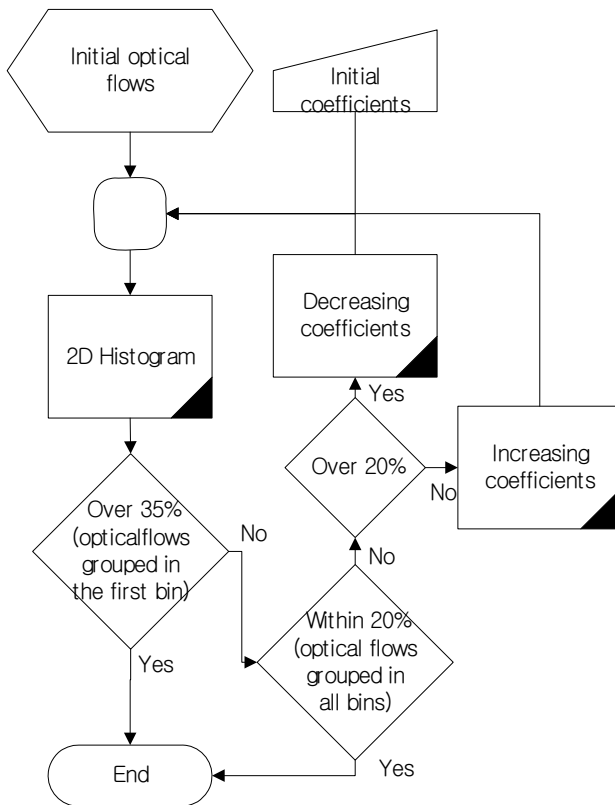


Fig. 2. Flow chart of the proposed algorithm

3 Experimental Results

LMedS is a well-known algorithm for detecting outliers with fitting given models as real-time computation. However, for cases in which the outlier rate is over 45%, discarding outliers with LMedS is very difficult. For images taken with a camera having

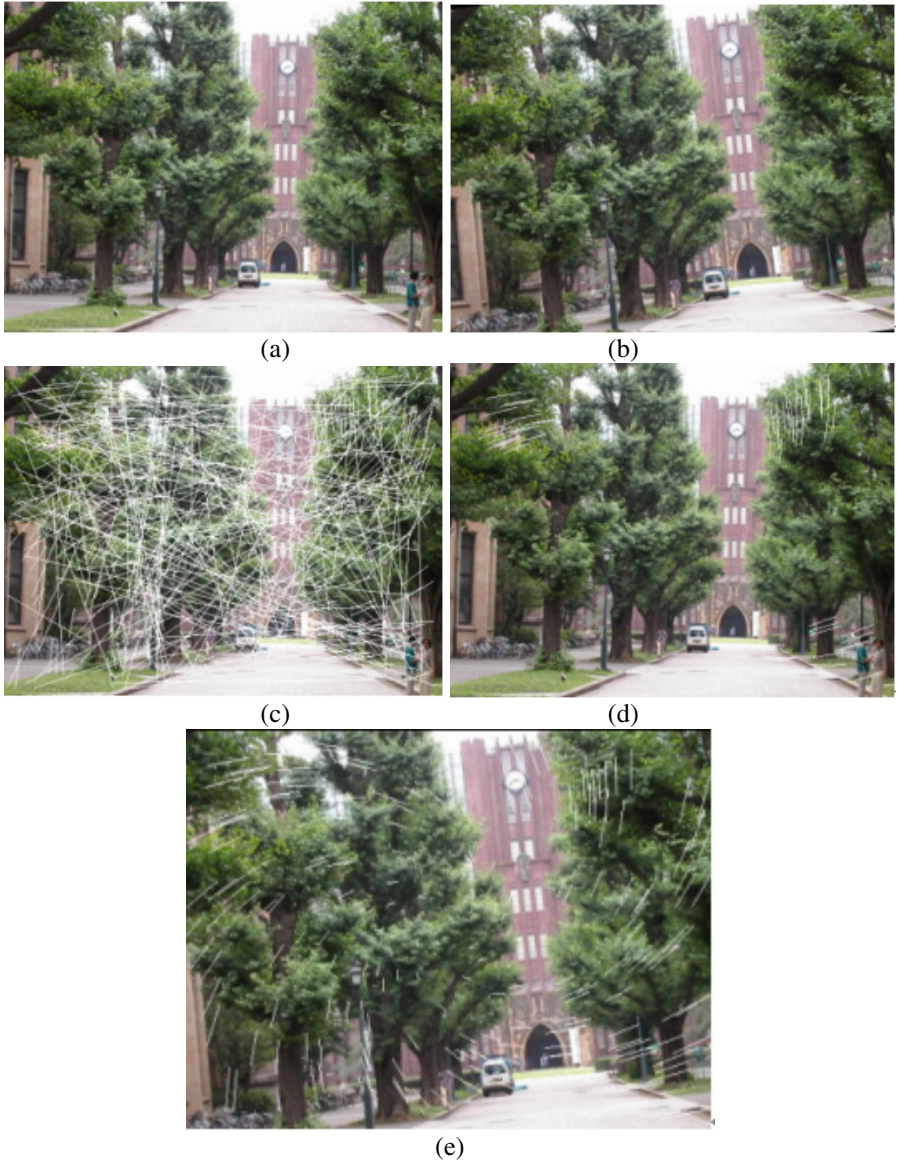


Fig. 10. Optical flow detection on the images taken while the camera rotates and zooms-in; (a) reference image frame. (b) subsequent image frame, (c) optical flows obtained with correlation-based matching, (d) selected optical flows using the proposed method, (e) optical flows detected by affine transformation with the locally grouped optical flows.

a zoom motions and itself optical axis motions or a camera with big translating motion, images capturing many similar textures or moving objects, and so on, the outlier rate will be increased over 60%. To demonstrate that the proposed method can robustly detect outliers, this paper performed experiments using a moving camera, as

shown in Fig. 10. Using a SUSAN operator, we extracted over one hundred feature points from each respective input image. The mask size of correlation for matching the extracted feature points was 11 pixels. When correlation scores between the feature points extracted from two images are over 0.8, the pairs are set initial optical flows.

Two images, as shown in Figures 10(a) and (b) were taken with a camera with zoom-in motion and itself optical axis motions. As can be seen, the detected optical flows are very complex and irregular. The results of applying the proposed method to the detected optical flows shown in Figs. 10(c) are shown in Figs. 10(d). The figures clearly show that locally similar optical flows are grouped. Optical flows shown in Figs. 10(e) were detected by affine transformation with the locally grouped optical flows.

Table 1. Robustness comparison between the proposed and the LMedS algorithms

Outlier / inlier		Inlier1	Outlier1	$\frac{\text{Inlier1}}{\text{Inlier1} + \text{Outlier1}}$	Estimation
0.6	LMedS	31	2	0.92	Good
	Proposed	52	2	0.96	Good
0.8	LMedS	18	3	0.82	Good
	Proposed	51	1	0.97	Good
1.0	LMedS	14	8	0.64	Middle
	Proposed	55	2	0.96	Good
1.2	LMedS	13	10	0.56	Bad
	Proposed	51	1	0.971	Good
1.4	LMedS	×	×	×	×
	Proposed	53	3	0.93	Good
2.0	LMedS	×	×	×	×
	Proposed	54	5	0.90	Good
2.8	LMedS	×	×	×	×
	Proposed	50	20	0.714	Good
3.6	LMedS	×	×	×	×
	Proposed	45	29	0.606	Middle
4.0	LMedS	×	×	×	×
	Proposed	38	41	0.48	Bad

Table 1 shows a comparison between the proposed method and the LMedS estimator applying optical flows mixed the initially detected, as shown in Fig. 10(c), and randomly generated optical flows. Even when generating the same number of optical flows randomly, according to position of the generated optical flows, the results are only slightly changed. To compare results obtained with the proposed method and with the LMedS estimator, while avoiding this influence, we used the average result of one hundred experiments. Table 1 shows that the proposed method can robustly detect inliers, even if there are four times as many outliers as inliers. In contrast, the LMedS estimator could not detect inliers, even though the times is less than 1.0.

5 Conclusions

We have proposed an algorithm that can robustly discard outliers using two-dimensional histogram with variable resolution, even if outlier rate in image matching is over 85%. The adopted two-dimensional histogram is used to group similar optical flows. By changing the resolution of the two-dimensional histogram, the proposed method limits the amount of grouped optical flows as a percentage of all optical flows to within the range of 10%-20%. Our experiments demonstrated that the proposed method can robustly detect inliers among optical flows, even among a high numbers of outliers, and it can be applied to moving objects tracked by a moving camera.

We expect that computation costs for RANSAC will be noticeably lower when using inliers detected by the proposed method for epipolar geometry. In addition, the proposed method can be applied to a moving robot equipped with CCD camera for real-time detection of moving objects.

Acknowledgement

This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment), (IITA-2005-C1090-0502-0023)

References

1. J. K. Kearney, W. B. Thompson, and D. L. Boley, "Optical flow estimation: An error analysis of Gradient-based methods with local optimization", *IEEE Tr. on PAMI*, Vol. 9, No. 2, pp. 229-244, 1987.
2. E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion", *Journal of Optical Society of America*, Vol. 2, No. 2, pp. 284-299, 1985.
3. S. M. Smith and J. M. Brady, "Real-Time Motion Segmentation and Shape Tracking", *IEEE Tr. on PAMI*, Vol. 17, No. 8, 1995.
4. B. K. P. Horn and B. G. Schunck, "Determining Optical Flow ", *Artificial Intelligence 1981*, pp. 185-203.
5. G. Harris, "Determination of Ego-Motion From Matched Points", *Proc. Alvey Vision Conf.*, pp. 189-192, Cambridge UK, 1987.
6. L. Kitchen and A. Rosenfeld, "Gray Level Corner Detection", *Pattern Recognition Letters*, pp. 95-102, 1982.
7. Farzin Mokhtarian and Riku Suomela, "Robust Image Corner Detection Through Curvature Scale Space", *IEEE Tr. on PAMI*, Vol. 12, 1998.
8. S. M. Smith and J. M. Brady, "SUSAN - a new approach to low level image processing", *In IJCV*, Vol. 23, No. 1, pp. 45-78, 1997.
9. R. M. Haralick, H. Joo, C. Lee, X. Zhuang, V. Vaidya, and M. Kim, "Pose estimation from corresponding point data", *IEEE Tr. on SMC*, Vol. 19, No. 6, pp. 1426-1446, Nov. 1989.
10. P. J. Rousseeuw, "Least median of squares regression", *Journal of American Statistics Association*, Vol. 79, pp. 871-880, 1984.

11. M.A. Fischler and R.C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", *Comm. Of the ACM*, Vol. 24, pp. 381-395, 1981.
12. C.K. Tang, G. Medioni, and M.S. Lee, "N-Dimensional Tensor Voting and Application to Epipolar Geometry Estimation", *IEEE Tr. on PAMI*, Vol. 23, No. 8, 2001.
13. R. Hartly, "In defense of the eight-points algorithm", *IEEE Tr. on PAMI*, Vol. 19, No. 6, pp. 580-593, 1997.

Iris Recognition: Localization, Segmentation and Feature Extraction Based on Gabor Transform

Mohammadreza Noruzi¹, Mansour Vafadoost¹, and M. Shahram Moin²

¹Biomedical Engineering Faculty, Amirkabir University of Technology, Tehran, Iran
{noruzi, vmansur}@cic.aut.ac.ir

²Multimedia Dept., IT Research Faculty, Iran Telecom. Research Center, Tehran, Iran
moin@itrc.ac.ir

Abstract. Iris recognition is one of the best methods in the biometric field. It includes two main processes: “Iris localization and segmentation” and “Feature extraction and coding”. We have introduced a new method based on Gabor transform for localization and segmentation of iris in eye image and also have used it to implement an Iris Recognition system. By applying the Gabor transform to an eye image, some constant templates are extracted related to the borders of pupil and iris. These features are robust and almost easy to use. There is no restriction and no tuning parameter in algorithm. The algorithm is extremely robust to the eyelids and eyelashes occlusions. To evaluate the segmentation method, we have also developed a gradient based method. The results of experimentations show that our proposed algorithm works better than the gradient based algorithm. The results of our recognition system are also noticeable. The low FRR and FAR values justify the results of segmentation method. We have also applied different Gabor Wavelet filters for feature extraction. The observations show that the threshold used to discriminate feature vectors is highly dependant on the orientation, scale and parameters of the corresponding Gabor Wavelet Transform.

1 Introduction

Biometric features are unique and dependant on personal genetic. Iris is the most attractive feature in human authentication field and also is the most accurate after the DNA. The authentication by iris can even distinguish between twins since the texture of iris is affected by the environment that the person has grown in it. The error rate of authentication by iris is about one per million. This low error rate makes it very suitable for highly secure access control applications [1, 2].

Iris recognition includes two steps: at first step the iris area must be extracted and then the feature vector must be constructed from the texture of segmented iris. The judgment is done by comparing the feature vectors and extracting the similarity between them.

The main work in this area is done by Daugman. He developed a system and introduced a method for each of the above mentioned steps. He used gradient of image for iris segmentation and Gabor Wavelet for feature extraction and introduced a constant threshold for discriminating the feature vectors [2].

In this paper we introduce a new method for iris segmentation with two characteristics: first, it can find the position of iris automatically and second, it uses Gabor transform as a more robust feature for finding the borders of iris. There is no restriction on the position of eye and no tuning parameter in algorithm.

An Iris recognition system has also been implemented. Gabor wavelet filters have been used for feature extraction. The similarities between feature vectors have been analyzed and the results show that by changing the parameters of the Gabor filter, the discriminating threshold must be changed and tuned. In other words, it is possible to obtain similar results using different filters with different thresholds.

The remaining of this paper is organized as follows. In section 2 we will present a brief introduction to Gabor transforms. The proposed algorithm will be explained in section 3. It is divided into segmentation and feature extraction processes. The segmentation algorithm has 3 steps: Finding the position of pupil, Exaction of the interior border of iris, Exaction of the exterior border of iris. There is a novel idea in each step. Section 4 is dedicated to feature extraction and comparison. To evaluate the proposed method, we have developed a gradient based segmentation method. The results of experimentations are presented in section 5.

2 Gabor Transform

Gabor transform is a Time-Frequency transform. In two dimensions it is defined as follows [2]:

$$G(x, y) = e^{-\pi[(x-x_0)^2/\alpha^2 + (y-y_0)^2/\beta^2]} \cdot e^{-2\pi i[u_0(x-x_0) + v_0(y-y_0)]} \quad (1)$$

Where $G(x, y)$ depicts a filter with center (x_0, y_0) , $i = \sqrt{-1}$ is a constant, $\omega_0 = \sqrt{u_0^2 + v_0^2}$ is modulation frequency, $\theta_0 = \arctan(v_0/u_0)$ is angle of modulation and α, β are the effective lengths of filter in x and y axes respectively.

It can be shown that the Gabor transform is a Short Time Fourier Transform with a Gaussian window and a modulation in frequency domain, therefore some of its characteristics are similar to its parent. By applying the Gabor transform to an image, the frequency contents of windowed image in direction θ_0 will be extracted. There is a reverse relation between the time resolution and frequency resolution, the wider windows in time domain will extract more accurate frequency contents, but the position of these contents will be more ambiguous [3].

The suitable amounts for α, β are depend on the resolution of input image and its texture. For a given filter if the input image has higher resolution, then thinner borders will have greater affects on the filter output. This transform not only can extract the borders but also can illuminate the expansion angles of these borders. This interesting property of Gabor filter has made it a powerful tool for feature extraction [2, 3].

This filter has real and imaginary parts with even and odd symmetry, respectively. The even part has nonzero mean values in general. We have found that removing this mean value improves the results. This process changes the real part of Gabor filter to

a bandpass filter. As a result, the background illumination of image will have negligible effect on filter output. This filter is different from the Stretched-Gabor filter presented by Heitger et al. [4].

By applying a Gabor filter in different orientations and resolutions the Gabor Wavelet transform is formed. This wavelet is generated from a fixed Gabor elementary function by dilation, translation and rotation [5].

3 Iris Localization and Segmentation

In this section we propose a new for localization and segmentation the iris. The proposed method resembles to the gradient method in some extent, but it using Gabor coefficients instead of gradient vector. Therefore the well known gradient method will be explained at first and then the proposed method will be explained. We have also implanted a gradient based method and have compared its results with results of the proposed method.

The segmentation of iris is based on two assumptions: iris border is circular and iris and pupil are concentric. The methods developed by other researchers mostly have a pre-assumption about the position of pupil in input image. They mostly assume that the position of the center of pupil is approximately known. For examples they suppose that the eye image is located in center of image frame, thus the pupil center is close to the input image center.

If the above assumption is not satisfied an extra pre-processing must be applied to the input image, for estimating the center of pupil.

After approximating the center of pupil, some points around this center are chosen as candidates for exact pupil center location and then, a feature will be computed for every candidate point. The winner is the candidate with the highest feature.

This feature is computed as follows: it is obvious that the border between the pupil and iris is circular and the highest contrast between two areas belongs to their border. To find this border, some circles with different radius will be drawn from each candidate point and the values of image gradients on circle's perimeter are summed. The circle that is exactly fitted over the pupil border reaches the maximum value. The same method is repeated to find the exterior border of iris. This technique has been used in many researches [6]-[10].

Although the basics of our method are similar to those of the gradient method, but it does not use gradient or any derivation operator. It is completely based on Gabor transform. Our approach has 3 steps: estimation of the pupil center, extraction of the interior border of iris and extraction of the exterior border of iris.

As mentioned before, we have implemented another method based on gradient to evaluate the results of Gabor based method. Our gradient based method is slightly different from Daugman's method. In Daugman's method the radial gradient is used, but our gradient based method is based on horizontal gradient, i.e. it searches for vertical edges. It is obvious that our gradient based algorithm is much faster than Daugman's method, because it calculates the gradient of image once, in contrary to the Daugman's method which calculates the radial gradient for each circle separately.

3.1 Estimation of the Pupil Center

The position of pupil is estimated with good approximation by using the phase of Gabor coefficients of filtered image. There is no essential restriction on input image and no need for pre-filtering. By applying a suitable Gabor filter to an image in horizontal direction, the pupil will produce a constant template including 2 half circles (Fig. 1).

The results have shown that the effective window length must be greater than the half size of pupil and shorter than the size of pupil.

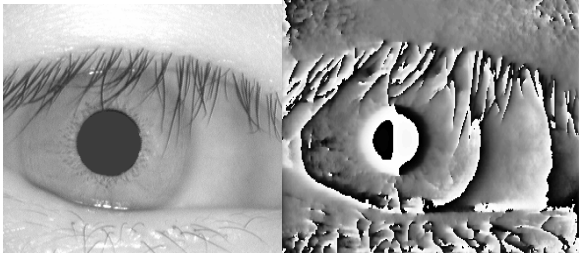


Fig. 1. The image of eye and the phase of its Gabor transform

A simple constant template has been used in our method to estimate the position of pupil. This template is driven directly from one of the images in the database and its size is equal to the size of the iris in source file. This template is shown in Fig. 2.



Fig. 2. Template used for estimating the pupil position

The estimation is done by applying the Normalized Cross Correlation [11] between the template and the phase of Gabor coefficients of transformed image.

It is noticeable that the algorithm has low sensitivity to the size of template. The size of template must be near the size of pupil. The practical results have shown that 50 percent change in the size of template will not affect on results. In other words, the algorithm has low sensitivity to the changes in the size of pupil, thus it is possible to find the position of pupils by using a constant template and a constant Gabor filter. We have used a constant Gabor filter and a constant template for all of the images in database.

To evaluate the proposed method for estimation of the pupil position, we have also implemented an alternative method. The pupil is like a dark circle in image. Therefore we define a new template like a dark circle and compute the cross correlation between the new template and the raw image. This method can also estimate the position of

pupil but the experimental results have shown that this new method is not as good as the Gabor based method. The mean error of pupil center estimation in gradient based method is twice the mean error of Gabor based method.

After the estimation of pupil center the interior and exterior borders of iris must be extracted. The next two sections are dedicated to this problem.

3.2 Extraction of Interior Border of Iris and Its Center

Suppose that the coordinates of the center of pupil are estimated. Obviously, the exact location of the center of the circle that is fitted on interior border of iris is close to the estimated center. Thus, some points around the estimated center are chosen as candidates for real location of center. Then, related to each candidate point, a series of circles with different radius are drawn in transformed image.

In the next step, the phase values of filtered image are computed on circumference of each circle and are summed over 360 degrees. As it's shown in Fig. 1, the left and right half circle of interior border of iris have opposite phase values. Therefore the sign of these values must be considered in the summation. Finally the accumulated value is divided to the radius of related circle. This normalized feature is the criterion for selecting the best fitted circle: the circle with highest radius and highest feature value is chosen.

3.3 Extraction of Exterior Border of Iris and Its Center

The exterior boundary of iris is determined in a similar way which is used for interior boundary. The upper and lower areas of iris are sometimes covered by eyelids and eyelashes. To solve this problem, the accumulated phase value feature is not computed in 360 degrees. As shown in Fig. 3, the eye image is divided into four 90-degree areas and features are computed on left and right quarters.

As the interior and exterior borders of iris are not concentric [2], some points around the exact center of interior border are chosen as candidate for exact center of exterior border. Then, like the previous section, some circles with variable radius are drawn in transformed image and the feature values are computed for every circle. As the radius of exterior border is greater than the radius of interior border and also for preventing possible mistakes, the radius of each circle must be greater than the radius of interior circle which is computed previously. The rest of algorithm is like the one used for interior border extraction.

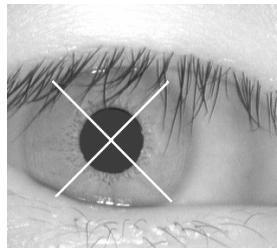


Fig. 3. Because of eyelids and eyelashes, the exterior border computations are done in left and right quarters

4 Feature Vector Extraction and Comparison

As mentioned before, the upper and lower parts of iris are usually occluded by eyelashes and eyelids. Therefore, the feature vector is extracted from the lateral part of iris image [2]. We divide the iris image to the 4 equal quarters as shown in Fig. 3. The left and right quarters are then mapped to a rectangular image so called dimensionless image (Fig. 4). The size of the mapped images is 256x64 and is identical for all of the images in database. The details of the mapping method can be found in [2].

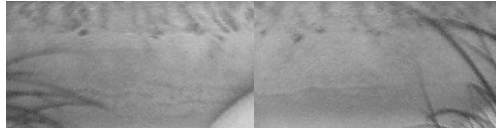


Fig. 4. An example of mapping the left and right parts of a iris image to a rectangular image

The reason of this mapping is that the most of fast and conventional techniques for feature extraction are implemented in Cartesian coordinate system.

The feature extraction is done by applying the Gabor transform in different resolutions and orientations.

It has been shown that the most amounts of information are in the phase of Gabor transform [5]. This characteristic has been used in iris recognition technique [2].

After applying the Gabor filters to the rectangular mapped image, the amplitude of output is put aside and just the sign of real part and imaginary part of the output is used as discriminating features. The final feature vector is a series of 1 and 0 bits. This reduces the complexity of feature vector and accelerates the comparison procedure. Although some of information are eliminated, but the remained information are enough for constructing a reliable system.

For comparing the feature vectors the simplest way is hamming distance which is done by applying Exclusive-OR operator on each pair of feature vectors [2]. If the distance between two feature vectors is lower than a threshold value then they belong to one person.

5 Experimental Results

We have tested the system using different Gabor Wavelet filters and obtained the following results:

(1)- Ma in [12] has reported that the most information is in rotational direction in iris image and implemented a system based on this characteristic. Our results showed that this characteristic is very powerful but one direction is not enough for a system that is based on Gabor transform.

(2)- The threshold value that is used for discriminating the feature vectors is dependant on orientation, scale and parameters of Gabor transform. It was observed that it is possible to implement different systems with different thresholds but with the same identification rate.

(3)- The parts of iris that are near the interior border have much more information than the other parts. It was observed that using 75 percents of iris image reduces the error rate 50 percents and improves the results. This characteristic is also used in the system that has been developed by Ma [12]. The final image is shown in fig 5.



Fig. 5. An example of truncated mapped image of iris

We implemented the system in identification mode. In this case the feature vector of each image is compared with the others feature vectors.

The CASIA¹ Iris Database has been used in our experiments. It contains 756 eye images from 108 persons. The resolution of images is 320x280. This database has also been used by other researchers and it has been reported explicitly that some errors of authentication are resulted from miss localization of irises [13, 14]². The database includes 108 persons and 7 pictures for each person. These 7 pictures have taken in two sessions. Therefore we suppose that a person has enrolled in one session and is identified in another session, in other words each image in a session of a person is compared with the all of the images in other sessions.

A program in MATLAB client and on a 2400 MHz PC has been written to simulate our proposed method. The total run time for each segmentation process is 7 seconds.

The segmentation method has applied to each image in database. Some of the results are shown in Fig. 6. Presence of eyelids and eyelashes in eye image are the most important problem for iris localization and segmentation. Images in Fig. 6 are the samples containing this trouble clearly.

For comparison, a gradient based method has also been implemented. The cross correlation between a dark circle and iris image is used to estimate the pupil position in gradient based method. The elapsed time for the gradient base method has been 5 seconds for each image in MATLAB client. It is noticeable that the performance of our implemented gradient based algorithm seems to be better than Daugman's method in CASIA database [14].

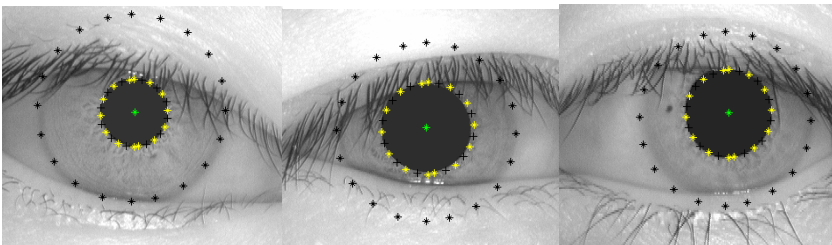


Fig. 6. Extracted borders for some samples in database

¹ Chinese Academy of Sciences Institute of Automation.

² The results of these articles are qualitative and do not report any percentage to compare with the results of the proposed algorithm.

The results have been analyzed by experts. The acceptance and rejection criteria are the accuracy of fitness between the borders which have been extracted by the program and the experts. Averaging the results shows that the experts have judged using following rule: if the difference between positions of two borders is below 7 percents of the corresponding circle, then the segmentation is assumed successful. The results of segmentations are shown in Table 1.

Table 1. Result of two implemented segmentation methods judged by experts

	Elapsed Time For Each Image (Seconds)	Interior Border Extraction Error %	Exterior Border Extraction Error %
Gradient Based Method	5	3	3.5
Gabor Based Method	7	1.5	2.5

The results of identification system justify the results of proposed segmentation method. Many different Gabor filter have been used in identification process. For example in an implementation with the threshold value equal to 0.32, the FAR³ and FRR⁴ were obtained equal to zero and in another implementation using different Gabor Wavelet filter with threshold value equal to 0.36, the same results were obtained again. In other words, it was observed that the shape of statistical distributions of hamming distances is almost constant, but it is shifted slightly as a result of changing the parameters of Gabor Wavelet transform.

The results show that the desirable length for a feature vector in a practical system varies between 400 to 800 bytes.

6 Conclusions

A novel method has been introduces based on Gabor transform was introduced for iris segmentation and localization. Two important characteristics of method are: (1) automatic localization and segmentation of iris with good accuracy and (2) noticeable speed of processing.

A gradient based method has also been implanted to evaluate the performance of proposed method. The implemented gradient based method is slightly different from the Daugman's method due to use of horizontal gradient instead of radial gradient. Since the horizontal gradient is computed once, our gradient based method is faster than Daugman's method. It is noticeable that the performance of gradient based algorithm is better than Daugman's method in CASIA database.

³ False Acceptance Rate

⁴ False Rejection Rate

An important question remains concerning about the method for pupil center estimation: although the implemented algorithm works correctly for the pictures in our database, how we can be assured that it works well for a new sample? In other words, how good a constant template and a constant filter work for every eye image? Of course, we can not guaranty that. The solution for this problem is multi resolution pattern matching: expanding the current method by using multiple filters and templates resolutions, for example 3 values for each of them. As shown in Fig. 1, it is obvious that the unique pattern of pupil will be revealed at least in one of the transformed images. In the case of using different resolution patterns and filters, the final result can be the pair with the maximum correlation.

Excepting the speed, the performance of Gabor based segmentation algorithm is better than gradient based method. Although the Gabor based algorithm is 40 percents slower than the gradient based, but it can be still considered fast enough for practical applications.

The proposed algorithm for segmentation has no tuning parameter and no threshold. The algorithm is not sensitive to eyelashes. The used Gabor transform is band pass and thus the back ground illumination has the lowest effect on the results. These remarkable features make the algorithm reliable and practical for iris recognition applications.

Acknowledgment

The authors would like to thank Chinese Academy of Sciences Institute of Automation for preparing the CASIA database.

References

1. Wilds, R.P.: Iris recognition: An emerging biometric technology. Proceeding of IEEE, Vol. 85, No. 9. (1997) 1348-1363.
2. Daugman, J.G.: High confidence visual recognition of persons by a test of statistical independence. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 15, No. 11. (1993) 1148-1161.
3. Chen, D.: Joint Time-Frequency Analysis. Prentice Hall, (1996).
4. Heitger, F., Rosenthaler, L., von der Heydt, R., Peterhans, E., & Kubler, O.: Simulation of neural contour mechanisms: from simple to end-stopped cells. Vision Research, Vol. 32. (1992) 963-981.
5. MacLennan, B.: Gabor representation of spatiotemporal visual images. Tech. Report, CS-91-144. Comp. Science Dept., Univ. Tennessee, (1994).
6. Lee, T.: Image representation using 2D Gabor wavelets. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, (1996) 959-971.
7. Ma, L., Wang, Y., Tan, T.: Iris Recognition Based on Multichannel Gabor Filtering. Proceedings of ACCV, Vol. I. Australia, (2002) 279-283.
8. Ma, L., Wang, Y., Tan, T.: Iris Recognition Using Circular Symmetric Filters. IEEE International Conference on Pattern Recognition, Vol. II. pp. Canada, (2002) 414-417.
9. Tisse, C., Martin, L.: Person Identification Technique Using Human Iris Recognition. Proc. of Vision Interface. (2002) 294-299.

10. Sanchez-Reillo, R., Sanchez-Avila, C.: Iris Recognition with Low Template Size. Proc. of Audio and Video Based Biometric Person Authentication. (2001)324-329.
11. Haralick, R.M., Shapiro, L.G.: Computer and Robot Vision, Vol. II. Addison-Wesley, (1992) 316-317.
12. Ma, L.: Local Intensity Variation Analysis for Iris Recognition. Pattern Recognition, Vol. 37, No. 6. (2004) 1287-1298.
13. Ma, L., Tieniu, T.: Efficient Iris Recognition by Characterizing Key Local Variations. IEEE Trans. on Image Processing, Vol. 13, No. 6. (2004) 739- 750.
14. Ajdari-Rad, A., Safabakhsh, R.:Fast Iris and Pupil Localization and Eyelid Removal Using Gradient Vector Pairs and Certainty Factors. Proc. of Conf. Machine Vision and Image Processing. (2004) 82-91.

Optimal Edge Detection Using Perfect Sharpening of Ramp Edges

Eun Mi Kim¹, Cherl Soo Pakh², and Jong Gu Lee³

¹ Dept. of Computer Science, Howon Univ., Korea
ekim@mail.howon.ac.kr

² Dept. of Visual Optics and Optometry, Daebul Univ., Korea
pcs@mail.daebul.ac.kr

³ Dept. of Computer Science, Chonbuk Univ., Korea
jgleemail@chonbuk.ac.kr

Abstract. The strictly monotonic intensity profiles of ramp edges are mapped to a simple step function of intensity. Such a perfect sharpening of ramp edges can be used to define a differential operator nonlocally, in terms of which we can construct an edge detection algorithm to determine the edges of images effectively and adaptively to varying ramp widths.

1 Introduction

Since the image sensing devices have physical limits in resolution and the limits are not uniform in general the edge features of resultant images are manifested with widely varying widths of intensity ramp. Therefore, in order to detect and locate such edge features precisely in real images we have developed an algorithm by introducing a nonlocal differentiation of intensity profiles called *adaptive directional derivative* (ADD), which is evaluated independently of varying ramp widths[1]. Contrary to the usual conventional edge detectors, which employ the local differential operator such as the gradient together with different length scales in the inspection of intensity changes[2][3] or additional matched filters[4][5][6], the ramp edges are identified as strictly monotonic intervals of intensity profiles and these nonlocal features are described by the values of ADD's together with the widths of the intervals. As in the Gaussian model of edge template, where the profile of edge signal intensity is described by the integral of a Gaussian added to a fixed background, we have two independent parameters for the edge detector. The one is the amplitude of edge corresponding to the value of ADD, and the other is the variance of edge indicating the edge acuity and hence related to the ramp width. Therefore we can detect ramp edges by evaluating ADD regardless of the ramp width or length scale.

In this paper, we first review the edge detector employing the ADD and then, we modify the ADD by introducing an extrinsic map which transfers the strictly monotonic intensity profiles of ramps to optimal simple step functions and is referred to as *perfect sharpening map*. The step functions are determined to

minimize their mean square errors to the real intensity profiles and the values of directional derivative of this optimal step functions are assigned to the original intensity profiles to define the *modified adaptive directional derivative* (MADD). By using the MADD instead of the ADD, we have the exact location of edge pixels determined naturally as the positions of the optimal simple step mapped to by the perfect sharpening map without bothering extra procedure to find the position such as the *local center of directional derivative* (LCDD)[1], which is the position averaged out of the pixels within the ramp by the weight of their directional derivatives. Our edge detection algorithm employing the MADD in x and y directions is completed with an elaborate edge trimming procedure added in to avoid the double manifestations resultant from identifying edges in both of the above two directions. The performance of the algorithm beyond the magnification of images is illustrated by comparing the results to those from the Canny's edge detector.

2 Adaptive Directional Derivative for Ramp Edges

In actual images, edges are blurred to yield ramp-like intensity profile due to optical condition, sampling rate and other image acquisition imperfection. However, the criterion by the usual local operator such as gradient or directional derivative may omit or repeat identifying a ramp edge with gradually changing gray level. In order to detect ramp edges properly, we employ the ADD of 1-D intensity profile. The pixel on the image is represented as a 2-D position vector like \mathbf{p} . Then, the 1-D profile crossing the pixel \mathbf{p} in the θ -direction is described by a gray-level valued function $f(\mathbf{p} + q\mathbf{u}_\theta)$ defined on a part of the linear sequence $\{\mathbf{p} + q\mathbf{u}_\theta; q \in Z\}$ in terms of the θ -direction vector \mathbf{u}_θ such as

$$\mathbf{u}_{\pm x} \equiv \pm(1, 0), \quad \mathbf{u}_{\pm y} \equiv \pm(0, 1), \quad \mathbf{u}_{\pm+} \equiv \pm(1, 1), \quad \mathbf{u}_{\pm-} \equiv \pm(1, -1). \quad (1)$$

Then the directional derivative of intensity (gray level) in the θ -direction is defined as

$$v_\theta(\mathbf{p}) \equiv D_\theta f(\mathbf{p}) = f(\mathbf{p} + \mathbf{u}_\theta) - f(\mathbf{p}), \quad (2)$$

and extended nonlocally to define ADD in the corresponding direction as

$$\Delta_\theta(\mathbf{p}) \equiv [1 - \delta_{s_\theta(\mathbf{p}), s_\theta(\mathbf{p} - \mathbf{u}_\theta)}] \sum_{k=0}^{\infty} \delta_{k+1, N_\theta(\mathbf{p}, k)} v_\theta(\mathbf{p} + k\mathbf{u}_\theta), \quad (3)$$

where

$$N_\theta(\mathbf{p}, k) \equiv \sum_{n=0}^k |s_\theta(\mathbf{p} + n\mathbf{u}_\theta)| \delta_{s_\theta(\mathbf{p}), s_\theta(\mathbf{p} + n\mathbf{u}_\theta)}, \quad (4)$$

$$s_\theta(\mathbf{p}) \equiv \begin{cases} +1, & v_\theta(\mathbf{p}) > 0 \\ 0, & v_\theta(\mathbf{p}) = 0 \\ -1, & v_\theta(\mathbf{p}) < 0 \end{cases} \quad (5)$$

and $\delta(\cdot, \cdot)$ is the Kronecker delta. According to this definition, $\Delta_\theta(\mathbf{p})$ has a nonvanishing value only when \mathbf{p} is the starting pixel of a strictly monotonic interval of the profile of f in the θ -direction. That is, if f starts to increase or decrease strictly at \mathbf{p} and ends at $\mathbf{p} + w\mathbf{u}_\theta$ from $f(\mathbf{p})$ to $f(\mathbf{p} + w\mathbf{u}_\theta)$ along the θ -direction then

$$\begin{aligned} \Delta_\theta(\mathbf{p}) &= \sum_{q=0}^{w-1} v_\theta(\mathbf{p} + q\mathbf{u}_\theta) \\ &= f(\mathbf{p} + w\mathbf{u}_\theta) - f(\mathbf{p}) \equiv A. \end{aligned} \tag{6}$$

According to the definitions, $v_\theta(\mathbf{p} + q\mathbf{u}_\theta) = -v_{-\theta}(\mathbf{p} + (q + 1)\mathbf{u}_\theta)$ and in the strictly monotonic interval $[\mathbf{p}, \mathbf{p} + w\mathbf{u}_\theta]$, $\Delta_\theta(\mathbf{p}) = -\Delta_{-\theta}(\mathbf{p} + w\mathbf{u}_\theta)$ is equal to the overall intensity change throughout the interval together with $\Delta_\theta(\mathbf{p} + q\mathbf{u}_\theta) = 0$ for $0 < q < w$. Hence we can notice that the ADD Δ_θ specifies the strictly monotonic change of intensity and the corresponding interval $[\mathbf{p}, \mathbf{p} + w\mathbf{u}_\theta]$ along some fixed θ -direction associated with a ramp edge. Thus we are available of a criterion for ramp edges in terms of Δ_θ with a threshold value T :

(CRE) *The absolute value of an adaptive directional derivative of gray level at \mathbf{p} is larger than T , i. e. $|\Delta_\theta(\mathbf{p})| \geq T$ for a θ -directions.*

With respect to a ramp edge where two regions on the image are bounded by the change of intensity A , we introduce the ϕ -direction which makes an angle of magnitude ϕ from the direction normal to the edge. We also introduce the corresponding direction vector \mathbf{u}_ϕ and ramp width w_ϕ , in terms of which the ramp profile in the ϕ -direction beginning at \mathbf{p} is described by a strictly monotonic gray level distribution over the interval $[\mathbf{p}, \mathbf{p} + w_\phi\mathbf{u}_\phi]$. \mathbf{u}_0 is normal and $\mathbf{u}_{\pm\pi/2}$ is tangential to the edge. If the edge is convex in the positive \mathbf{u}_0 direction, $\Delta_\phi(\mathbf{p}) = A$ all for $-\pi/2 \leq \phi \leq \pi/2$. If the edge is concave in the positive \mathbf{u}_0 direction, $\Delta_\phi(\mathbf{p}) = A$ only for a restricted range of angle near $\phi = 0$ and $\Delta_\phi(\mathbf{p}) = 0$ definitely at $\phi = \pm\pi/2$. Therefore, in order to detect the intensity changes associated with all the ramp edges by estimating the ADD Δ_θ in general, we should employ at least two of them Δ_{θ_1} and Δ_{θ_2} with different fixed directions θ_1 and θ_2 making an right angle.

Employing the criterion (CRE), the local maximal length interval $[\mathbf{p}, \mathbf{p} + w\mathbf{u}_\theta]$ of strictly monotonic gray level distribution along the θ -directions satisfying $|f(\mathbf{p} + w\mathbf{u}_\theta) - f(\mathbf{p})| \geq T$ is identified as the width of a ramp edge, and one of $w + 1$ pixels in the interval is determined as the edge pixel. The detection of the edges of all directions can be performed by using the ADD $\Delta_{\theta_1}(\mathbf{p})$ or $\Delta_{\theta_2}(\mathbf{p})$ of two orthogonal directions and we can use the two fixed directions $\theta_1 = x$ and $\theta_2 = y$ in practice. Here, we note that the criterion (CRE) can detect the edge regardless of spatial blurring or scaling which are concerned with the edge width w . However, we need a natural procedure in addition in order to determine the precise location of a single edge pixel out of the ramp interval $[\mathbf{p}, \mathbf{p} + w\mathbf{u}_\theta]$ such as the LCDD.

3 Modified Adaptive Directional Derivative and Perfect Sharpening of Ramp Edges

With respect to a ramp edge, which is identified as a strictly monotonic interval in the 1-D intensity profile in some direction on the image, we adopt an ideal step edge as the result of its perfect sharpening. That is, each (ramp) edge is associated with a unique step edge located at a certain pixel within the ramp with the same amplitude (i. e. overall intensity change throughout the whole ramp). If we are provided with a relevant rule to locate a unique step edge associated with the ramp edge, we can assign the directional derivative of the intensity of the associated ideal step edge to the ramp edge so that we define it as the MADD in the corresponding direction on the image where the 1-D intensity profile is identified as this strictly monotonic function. The strictly monotonic 1-D intensity profile $f(q)$ representing a ramp edge over the interval $[0, w]$ of q with $f(0) = B$ and $f(w) = A + B$, for example, is associated with a ideal step edge whose intensity profile is described by

$$f_s(q) = \begin{cases} B, & 0 \leq q \leq q_E \\ A + B, & q_E < q \leq w \end{cases} \tag{7}$$

for the relevant q_E within the interval $[0, w]$ (Fig. 1).

In the pixel space manifestation, the directional derivative of f_s is determined as

$$\frac{\partial f_s}{\partial q}(q) \equiv f_s(q + 1) - f_s(q) = \begin{cases} A, & q = [q_E] \\ 0, & q \neq [q_E] \end{cases} \tag{8}$$

since q should be integer-valued. If the map $S : f \rightarrow f_s$, which assigns a unique step function f_s of the amplitude A to each strictly monotonic function f with the same amplitude and performs the perfect sharpening, is well defined, we can define the MADD of f as

$$\tilde{\Delta}_f(q) \equiv \frac{\partial f_s}{\partial q}(q) \tag{9}$$

for every $q \in [0, w]$. In order to define the map S completely, we still have to determine the relevant location q_E of the step. The relevance can be granted

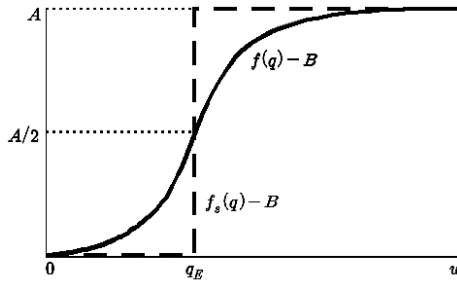


Fig. 1. Perfect sharpening of the ramp intensity profile

by minimizing the mean square error associated with the map S , $MSE(S)$. Provided the intensity function is *independent and identically distributed*, the mean square error reduced to

$$\begin{aligned}
 MSE(S) &= \int_0^w P(f_s; q)[f_s(q) - f(q)]^2 dq \\
 &= \int_0^w [f_s(q) - f(q)]^2 dq \\
 &= \int_0^{q_E} [B - f(q)]^2 dq + \int_{q_E}^w [A + B - f(q)]^2 dq,
 \end{aligned} \tag{10}$$

together with a uniform density $P(f_s; q) = 1$. The value of q_E minimizing $MSE(S)$ satisfies

$$\frac{d}{dq_E} MSE(S) = [B - f(q_E)]^2 - [A + B - f(q_E)]^2 = 0, \tag{11}$$

which dictates the step locating condition (SLC) at $q = q_E$;

$$\text{(SLC)} \quad f(q_E) - B = \frac{A}{2}.$$

With respect to the strictly monotonic function, such a position $q = q_E$ is uniquely determined. (Fig. 1) When f itself is a step function or constant, the minimum of $MSE(S)$ is trivially accomplished by $f_s = f$ without bothering the condition SLC and the perfect sharpening map S is well defined.

In the pixel space, the 1-D intensity profile $f(\mathbf{p} + q\mathbf{u}_\theta)$ with $q \in Z$ in the θ -direction of an image can be divided into a sequence of strictly monotonic or constant intensity intervals. With respect to each of these intervals, the profile can be mapped to the step (or constant) function by the map S , which can be naturally extended over the whole of intensity profile to obtain $f_s(\mathbf{p} + q\mathbf{u}_\theta)$ and also the MADD of $f(\mathbf{p} + q\mathbf{u}_\theta)$ can be defined as

$$\tilde{\Delta}_\theta(\mathbf{p} + q\mathbf{u}_\theta) = \frac{\partial f_s}{\partial q}(\mathbf{p} + q\mathbf{u}_\theta) \equiv f_s(\mathbf{p} + (q + 1)\mathbf{u}_\theta) - f_s(\mathbf{p} + q\mathbf{u}_\theta) \tag{12}$$

all over the whole intensity profile in the θ -direction. In the pixel space manifestation, the condition (SLC) is replaced by

$$\text{(ISLC)} \quad f(\mathbf{p} + [q_E]\mathbf{u}_\theta) - B \leq \frac{A}{2} \leq f(\mathbf{p} + ([q_E] + 1)\mathbf{u}_\theta) - B,$$

for the strictly increasing interval and

$$\text{(DSLCL)} \quad f(\mathbf{p} + [q_E]\mathbf{u}_\theta) - B \geq \frac{A}{2} > f(\mathbf{p} + ([q_E] + 1)\mathbf{u}_\theta) - B,$$

for the strictly decreasing interval. With respect to a strictly monotonic interval $[\mathbf{p}, \mathbf{p} + w\mathbf{u}_\theta]$ *non-extensible* (i. e. not extended to the larger one including it)

in the θ -direction with $f(\mathbf{p} + w\mathbf{u}_\theta) - f(\mathbf{p}) = A$, the MADD $\tilde{\Delta}_\theta(\mathbf{p} + q\mathbf{u}_\theta)$ ($q = 0, 1, 2, \dots, w-1$) is valued as $\tilde{\Delta}_\theta(\mathbf{p} + q\mathbf{u}_\theta) = A$ for $q = [q_E]$ and $\tilde{\Delta}_\theta(\mathbf{p} + q\mathbf{u}_\theta) = 0$ otherwise. We thus can find out that the MADD can detect the significant change A (of intensity) and its relevant location $q = [q_E]$ simultaneously. On substituting the MADD $\tilde{\Delta}_\theta$ for the ADD Δ_θ , the detection and location of edge pixels is accomplished simply by implementing the criterion on $\tilde{\Delta}_\theta$:

If the MADD of intensity in a θ -direction satisfies $|\tilde{\Delta}_\theta(\mathbf{p})| \geq T$ at a pixel \mathbf{p} , this pixel is an edge pixel.

With respect to the above strictly monotonic interval $[\mathbf{p}, \mathbf{p} + w\mathbf{u}_\theta]$, if $|A| \geq T$, $|\tilde{\Delta}_\theta(\mathbf{p} + [q_E]\mathbf{u}_\theta)| \geq T$ and $\mathbf{p} + [q_E]\mathbf{u}_\theta$ is an edge pixel according to this criterion. Since $\tilde{\Delta}_\theta(\mathbf{p} + [q_E]\mathbf{u}_\theta) = \Delta_\theta(\mathbf{p})$, the MADD's $\tilde{\Delta}_{\theta_1}$ and $\tilde{\Delta}_{\theta_2}$ in two orthogonal fixed directions can pick out all of the edges of an image as the ADD's do. While we can employ two MADD's $\tilde{\Delta}_x$ and $\tilde{\Delta}_y$ in order to detect all the edge of images, some parts of edges can be doubly identified both by $\tilde{\Delta}_x$ and $\tilde{\Delta}_y$. In particular when the edges are severely curved, the detected edge pixels would form hairy edge lines with a few cilia. Therefore, we implement some exclusive conditions that render only one of the two MADD's $\tilde{\Delta}_x$ and $\tilde{\Delta}_y$ applied case by case to obtain clear edge lines. The relevant condition is such that $\tilde{\Delta}_x$ is used to identify the edge only when $|v_x|$ is larger than $|v_y|$ and $\tilde{\Delta}_y$ is adopted otherwise. Hence we can describe our edge detection strategy as the following procedures:

P1. We scan the image along the x direction to find the strictly monotonic and nonextensible intervals such as $[\mathbf{p}, \mathbf{p} + w_x\mathbf{u}_x]$ where the overall change of intensity $|f(\mathbf{p} + w_x\mathbf{u}_x) - f(\mathbf{p})|$ i. e. $|\tilde{\Delta}_x(\mathbf{p} + [q_E]\mathbf{u}_x)|$ for $[q_E] \in [0, w_x - 1]$ satisfies the criterion $|\tilde{\Delta}_x(\mathbf{p} + [q_E]\mathbf{u}_x)| \geq T$.

P2. Locate an edge pixel at $\mathbf{p} + [q_E]\mathbf{u}_x$ within the interval where the condition (ISLC) or (DSLCL) with $\theta = x$ is satisfied only when $|v_x(\mathbf{p} + [q_E]\mathbf{u}_x)| \geq |v_y(\mathbf{p} + [q_E]\mathbf{u}_x)|$.

P3. Repeat the procedures with respect to the y direction to find the strictly monotonic and nonextensible intervals such as $[\mathbf{p}, \mathbf{p} + w_y\mathbf{u}_y]$ satisfying $|\tilde{\Delta}_y(\mathbf{p} + [q_E]\mathbf{u}_y)| \geq T$ for $[q_E] \in [0, w_y - 1]$.

P4. Locate an edge pixel at $\mathbf{p} + [q_E]\mathbf{u}_y$ within the interval where the condition (ISLC) or (DSLCL) with $\theta = y$ is satisfied only when $|v_y(\mathbf{p} + [q_E]\mathbf{u}_y)| \geq |v_x(\mathbf{p} + [q_E]\mathbf{u}_y)|$.

We can notice that these procedures can detect an edge regardless of spatial blurring or scaling which are concerned with the edge width parameter w_x and w_y , and extract the exact edge simply by deciding such pixels at $q = [q_E]$ that the half of the overall intensity change throughout the strictly monotonic and nonextensible interval occurs at $q = [q_E]$ or between $[q_E]$ and $[q_E + 1]$. Actually

magnifying the image by the scale κ makes no differences in the crucial conditions $|\tilde{\Delta}_x(\kappa\mathbf{p} + [\kappa q_E]\mathbf{u}_x)| \geq T$, $|v_x(\kappa\mathbf{p} + [\kappa q_E]\mathbf{u}_x)| \geq |v_y(\kappa\mathbf{p} + [\kappa q_E]\mathbf{u}_x)|$, etc. of this procedure except for changing to $\kappa\mathbf{p}$, κw_x , κw_y and $[\kappa q_E]$. Here, we note that $v_\theta(\kappa\mathbf{p} + \kappa q\mathbf{u}_\theta) = (1/\kappa)v_\theta(\mathbf{p} + q\mathbf{u}_\theta)$ while $\tilde{\Delta}_\theta(\kappa\mathbf{p} + \kappa q\mathbf{u}_\theta) = \tilde{\Delta}_\theta(\mathbf{p} + q\mathbf{u}_\theta)$ is left invariant. Therefore, the above procedure has the invariant performance beyond scaling contrary to the usual edge detectors employing local derivatives such as v_θ .

4 Algorithm and Applications

The entire edge detection procedure is performed by estimating $\tilde{\Delta}_x$ and $\tilde{\Delta}_y$ along the every vertical and horizontal lines of an image. With respect to the strictly increasing intervals in the x -direction, we apply the SIEDPS (strictly increasing edge detection by perfect sharpening) procedure to detect and locate an edge pixel, which is suspended at the end pixel \mathbf{p}_F of the line and specified in the following pseudo code:

Procedure SIEDPS

1. While $v_x(\mathbf{p}) > 0$ and $\mathbf{p} \neq \mathbf{p}_F$ with respect to the value $v_x(\mathbf{p})$ read from the successive scan pixel by pixel along the x -direction, do
 - (a) Put $\tilde{\Delta} = 0$, $FE = 0$ and $\mathbf{PE} = \mathbf{p}$ initially.
 - (b) Add $v_x(\mathbf{p})$ to $\tilde{\Delta}$ to obtain its new value.
 - (c) With respect to every $\tilde{\Delta}$, add $v_x(\mathbf{PE})$ to FE to obtain the new value of FE and substitute the next pixel for \mathbf{PE} to obtain the new value of $v_x(\mathbf{PE})$ until it satisfies $\tilde{\Delta}/2 < FE + v_x(\mathbf{PE})$.
2. When $v_x(\mathbf{p}) \leq 0$ or $\mathbf{p} = \mathbf{p}_F$ (i. e. at the end of the strictly increasing interval or line), write \mathbf{PE} as an edge pixel if $\tilde{\Delta} \geq T$ and $|v_x(\mathbf{PE})| \geq |v_y(\mathbf{PE})|$.

With respect to the strictly increasing intervals in the y -direction, we apply the above SIEDPS procedure with x -direction replaced by y -direction. With respect to strictly decreasing intervals, we substitute the SDEDPS (strictly decreasing edge detection by perfect sharpening) procedure, which is obtained simply by replacing $v_\theta(\mathbf{p})$ with $-v_\theta(\mathbf{p})$ in the SIEDPS procedure. In order to construct the algorithm for detecting edge pixels from scanning the whole of a single line in the $\theta = x$ or $\theta = y$ direction, we combine the SIEDPS and SDEDPS procedures as the sub-procedures for the pixels with $v_\theta(\mathbf{p}) > 0$ and $v_\theta(\mathbf{p}) < 0$ respectively, together with the bypassing procedure which produces no output for the pixels with $v_\theta(\mathbf{p}) = 0$. As the output of this algorithm for *edge detection by perfect sharpening of ramps* (EDPSR), we obtain the edge pixels identified with the pixels corresponding to the half of intensity change over edge ramps from the SIEDS or SDEDS procedures. Then, a complete edge detection algorithm of an image is constructed by the integration of EDPSR algorithm over all the lines of x and y direction of the image.

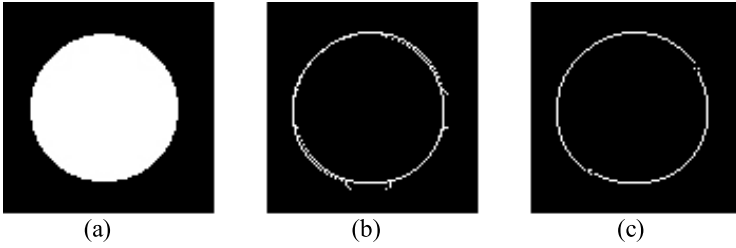


Fig. 2. Trimming the edge of disc

We first apply our new algorithm to a simple but illustrative image of circular disc. (Fig. 2a) In Fig. 2b, we have the result of edge detection by our former algorithm introduced in the section 2, where some hairs and doubling of edge line are observed implying double manifestation of the edge identified twice by the scans of x and y directions. We can find the hairs trimmed and doubled lines disappear in Fig. 2c, which is the result of applying our new algorithm. We can thus notice that the exclusive selection simply by comparing the absolute values $|v_x|$ and $|v_y|$ of directional derivatives out of the identifications by the values of MADD's $\tilde{\Delta}_x$ and $\tilde{\Delta}_y$ works relevantly.



Fig. 3. Comparison of the results of edge detection

We next apply the algorithm to a practical image, the picture of Lena. (Fig. 3a) The result Fig. 3c of our algorithm with the threshold $T = 30$ out of 256 gray levels is compared to the result Fig. 3b of Canny's in terms of Sobel mask with the equivalent lower threshold value 120. We can see equally (or even more) precise edges manifested successfully by the rather simple algorithm identifying strictly monotonic intervals out of the image. In Fig. 3b, the thick edge lines appear since every pixel over the lower threshold in the same ramp edge can be manifested and still further exhaustive thinning procedure should be applied to obtain precise edge location while every edge line has single pixel width representing precise edge location in Fig. 3c.

When the image is magnified κ times in length scale, the local derivatives in the pixel space such as Sobel operator have their values multiplied by $1/\kappa$

while our MADD $\tilde{\Delta}_\theta$ is invariant. Therefore, the Canny's edge detector may lose some edges after magnification unless the threshold is readjusted. However, our edge detector can remain equally effective beyond scaling. In Fig. 4a and 4b, two parts of the origin of Lena image Fig. 3a are magnified 4 times. The canny's edge detector yields the results of Fig. 4c and 4d and our edge detector yields Fig. 4e and 4f. We can observe that some edge lines manifested in Fig. 4e and 4f are lost in Fig. 4c and 4d in this magnification. That is, the edges are well identified by our edge detector still after some magnifications although some of them are lost by the edge detectors employing the local differential operators.

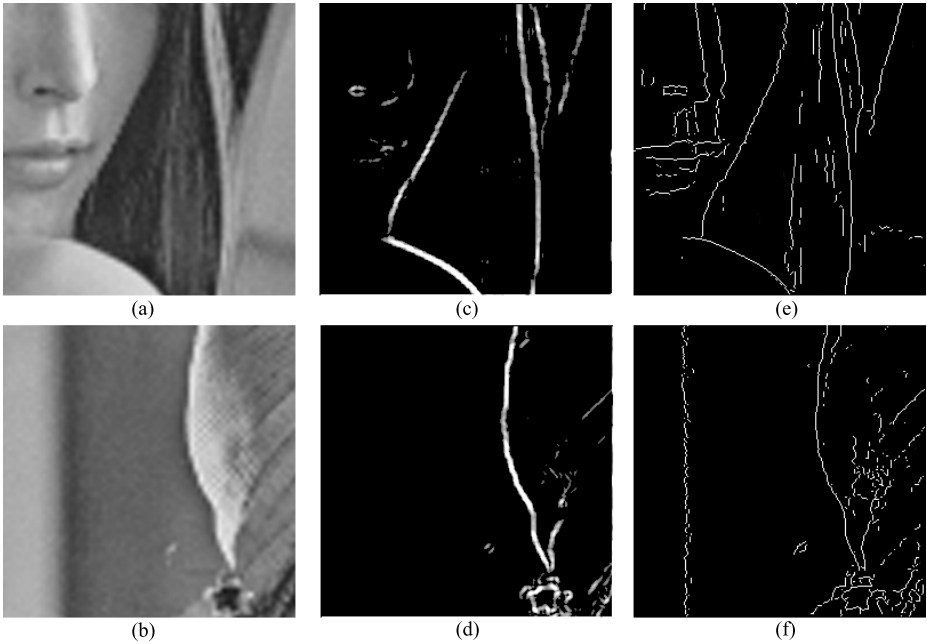


Fig. 4. Comparison of the results of edge detection by image magnification

5 Discussions

In order to verify the presence of edges and locate the precise edge pixels, we only have to estimate the MADD's $\tilde{\Delta}_x(\mathbf{p})$ and $\tilde{\Delta}_y(\mathbf{p})$ independently. Therefore, this algorithm seems to have relative simplicity of calculation compared to the usual ones employing the magnitude of gradient which include the floating point operation for the calculation of square root and needing extra thinning algorithm. Although the operators $\tilde{\Delta}_x(\mathbf{p})$ and $\tilde{\Delta}_y(\mathbf{p})$ are badly nonlocal, they can be estimated by single oneway scan of the local derivatives $v_x(\mathbf{p})$ and $v_y(\mathbf{p})$ as described in the procedures SIEDPS and so on, which causes no additional calculational expense. We also note that the additional calculations comparing

$|v_x|$ and $|v_y|$ in order to avoid the double manifestations of edges due to the double identifications by $\tilde{\Delta}_x$ and $\tilde{\Delta}_y$ cost not so much at all since the involved calculations are simple linear and very restrictive.

The false positive effect of the above double manifestation occurs when one of the strictly monotonic intervals associated with the MADD's $\tilde{\Delta}_x$ and $\tilde{\Delta}_y$ does not completely cross over the edge ramp between the two regions divided by it, which mostly implicates that the corresponding directional derivative has a smaller absolute value than the other. Therefore, such a false positive effect on edge detection can be suppressed efficiently by demanding that the corresponding directional derivative has a larger absolute value than the other, which has been not proved exactly.

Because of the nonuniformity in illumination, the gradual change of intensity over large area of a single surface may be recognized as a false edge in applying the MADD of intensity profile as the characteristic attribute of edge signal. In order to prevent this false positive effect in particular, we need the unsharp masking or flat-fielding procedure before applying the edge detector[7][8], or can restrict the width of strictly monotonic intervals to be detected.

Acknowledgements

This work was supported by the grants from the Howon University and the Daebul University.

References

1. E. M. Kim and C. S. Pakh, "Strict Monotone of an Edge Intensity Profile and a New Edge Detection Algorithm," LNCS 2690: Intelligent Data Engineering and Automated Learning, pp. 975-982 (2003).
2. D. Marr and E. Hildreth, "Theory of edge detection," Proc. R. Soc. London B207, 187-217 (1980).
3. J. Canny, "A computational approach to edge detection," IEEE Trans. Pattern Anal. Mach. Intell. PAMI-8, 679-698 (1986).
4. R. A. Boie and I. Cox, "Two dimensional optimum edge recognition using matched and Wiener filters for machine vision," in Proceedings of the IEEE First International Conference on Computer Vision (IEEE, New York, 1987), pp. 450-456.
5. R. J. Qian and T. S. Huang, "Optimal Edge Detection in Two-Dimensional Images." IEEE Trans. Image Processing, vol. 5, no. 7, pp. 1215-1220 (1996).
6. Z. Wang, K. R. Rao and J. Ben-Arie, "Optimal Ramp Edge Detection Using Expansion Matching." IEEE Trans. Pattern Anal. Machine Intell., vol. 18, no. 11, pp. 1586-1592 (1996).
7. J. R. Parker, "Algorithms for Image Processing and Computer Vision." Wiley (1997) and references therein.
8. M. Seul, L. O'Gorman and M. J. Sammon, "Practical Algorithms for Image Analysis." Cambridge U. Press, Cambridge (2000) and references therein.

Eye Tracking Using Neural Network and Mean-Shift

Eun Yi Kim¹ and Sin Kuk Kang²

¹ Department of Internet and Multimedia Engineering, Konkuk Univ., Korea
eykim@konkuk.ac.kr

² Department of Computer Engineering, Seoul National Univ., Korea
skkang@cglab.snu.ac.kr

Abstract. In this paper, an eye tracking method is presented using a neural network (NN) and mean-shift algorithm that can accurately detect and track user's eyes under the cluttered background. In the proposed method, to deal with the rigid head motion, the facial region is first obtained using skin-color model and connected-component analysis. Thereafter the eye regions are localized using neural network (NN)-based texture classifier that discriminates the facial region into eye class and non-eye class, which enables our method to accurately detect users' eyes even if they put on glasses. Once the eye region is localized, they are continuously and correctly tracking by mean-shift algorithm. To assess the validity of the proposed method, it is applied to the interface system using eye movement and is tested with a group of 25 users through playing a 'aligns games.' The results show that the system process more than 30 frames/sec on PC for the 320×240 size input image and supply a user-friendly and convenient access to a computer in real-time operation.

1 Introduction

Gesture-based interface have been considerable interests during the last decades that use human gestures to convey information such as input data or to control devices and applications such as computers, games, PDAs, etc.

Users create gestures by a static hand or body pose or a physical motion including eye blinks or head movements. Among them, human eye movements such as gaze direction and eye blinks have a high communicate value

Due to this, such an interface using eye movements has gained many attractions, so far many systems have been developed [1-4]. Then they can be classified into two major techniques: device-based techniques and video-based techniques. Device-based techniques use a glasses, head band, or cap with infrared/ultrasound emitters, which measure the user's motion using the changes in ultrasound waves or infrared reflector. In contrast to the device-based techniques using additional devices, the video-based techniques detects the user's face and facial features by processing the images or videos obtained via a camera. When compared with the device-based techniques, these are a non-intrusive and comfortable to users, and inexpensive communication device.

For practical use of these video-based systems, automatic detection and tracking of faces and eyes in real-life situation should be first supported. However, in most of the commercial systems, the initial eye or face position are manually given or some

conditions are used; the user initially clicks on the features to be tracked via a mouse in [5]; some systems require the user to blink only the eye for a seconds and then finds the eye regions via differencing the successive frames [6]. Moreover, they use the strong assumptions to make the problems more tractable. Some common assumptions are the images contain frontal facial view and the face has no facial hair or glasses.

To overcome abovementioned problems, this paper proposes a new eye tracking method using neural network (NN) and mean-shift that can automatically locate and track the accurate features under the cluttered background with no constraints. In our method, the facial region is first obtained using skin-color model and connected-component analysis. And then, the eye regions are localized by a NN-based texture classifier that discriminates each pixel in the extracted facial regions into the eye-class and non-class using the texture property. This enables us to accurately detect user's eye region even if they put on the glasses in the cluttered background. Once the eye regions are detected in the first frame, they are continuously tracked by a mean-shift.

To evaluate the proposed method, the method is applied to the interface to convey the user's command via his (her) eye movements. The interface system is tested with 25 numbers of people, and then the result shows that our method is robust to the time-varying illumination and less sensitive to the specula reflection of eyeglasses. Also, the results show that it can be efficiently and effectively used as the interface to provide a user-friendly and convenient communication device.

The remainder of this paper is organized as follows: Sections 2 describes the face extraction using skin-color model, then Section 3 provides a detailed description of the eye detection. The tracking process of the extracted eye regions is described in the next section. Thereafter, the experimental results are presented in Section 5, and Section 6 gives a final summary.

2 Skin-Color Based Face Extractor

Detecting pixels with a skin-color offers a reliable method for detecting face part. In the RGB space obtained by most video cameras, the RGB representation includes both color and brightness. As such, RGB is not necessarily the best color representation for detecting pixels with a skin-color [7-10].

The chromatic colors provide a skin-color representation by normalizing the RGB-value by its intensity. Therefore, the brightness can be removed by dividing the three components of a color pixel by the intensity. This space is known as chromatic color.

Fig. 1 shows the color distribution of human faces obtained from sixty test images in chromatic color space. Since the color distribution is clustered within a small area of the chromatic color space it can be approximated by a 2D Gaussian distribution. Therefore, the generalized skin-color model can be represented by 2D Gaussian distribution, $G(m, \Sigma^2)$, as follows,

$$m = (\bar{r}, \bar{g}), \quad \bar{r} = \frac{1}{N} \sum_{i=1}^N r_i, \quad \bar{g} = \frac{1}{N} \sum_{i=1}^N g_i, \quad \Sigma = \begin{bmatrix} \sigma_{rr} & \sigma_{rg} \\ \sigma_{gr} & \sigma_{gg} \end{bmatrix} \quad (1)$$

where \bar{r} and \bar{g} represent Gaussian means of r and g color distribution respectively, and Σ^2 represent covariance matrix of each distribution.

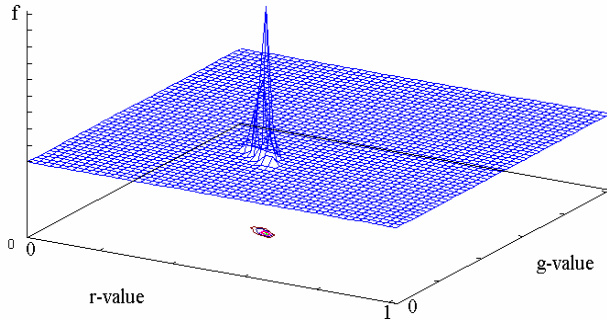


Fig. 1. The color distribution of human faces in chromatic color space

Once the skin-color model is created, the most straightforward way to locate the face is to match the skin-color model with the input image to identify facial regions. As such, each pixel in the original image is converted into chromatic color space, then compared with the distribution of the skin-color model. By thresholding the matched results, we obtain a binary image (see Fig. 2(b)). For the binary image, the connected component labeling is performed to remove the noise and small region. Then the facial regions are obtained by selecting the largest components with skin-colors, which is shown in Fig. 2(c).

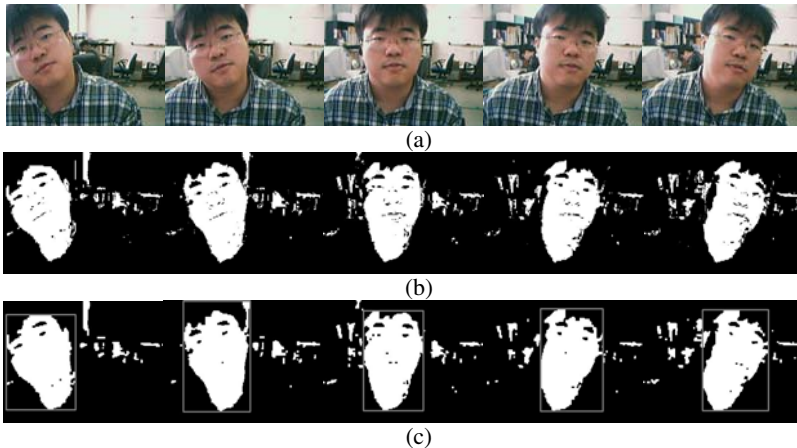


Fig. 2. Face detection results. (a) original color images, (b) skin-color detection results, (c) extracted facial regions

3 NN-Based Eye Detection

Our goal is to detect the eye in the facial region and track it through the whole sequence. Generally, the eye region has the following properties: 1) it has the high brightness contrast between white eye sclera and dark iris and pupil, along the texture

of the eyelid; 2) it has place in the upper of the facial region. These properties help reduce the complexity of the problem, and facilitate the discrimination between the eye regions from the whole face. Here, we use a neural network as a texture classifier to automatically discriminate the pixels of the facial regions into eye regions and non-eye ones in various environments. The network scans all the pixels in the upper facial region so as to classify them as eye or non-eye. The network receives the gray-scale value of a pixel and its neighboring pixel within a small window to classify the pixel as eye or non-eye. Then, the output of the network represents the class of the central pixel in the input window. A diagram of our eye detection scheme is shown in Fig. 3.

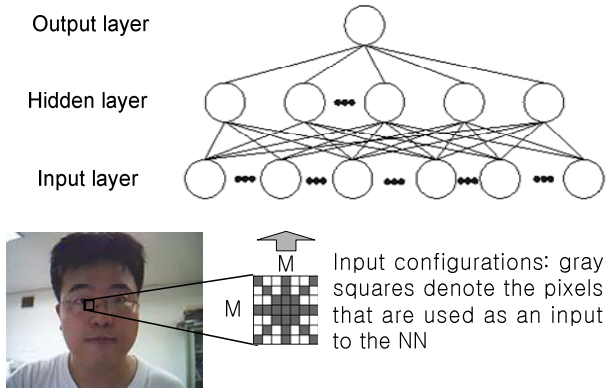


Fig. 3. A diagram of eye detection scheme

3.1 Texture Classification

We assume that the eye has a different texture from the facial region and is detectable. The simplest way to characterize the variability in a texture pattern is by noting the gray-level values of the raw pixels. This set of gray values becomes the feature set on which the classification is based. An important advantage of this approach is the speed with which images can be processed, as the features do not need to be calculated. However, the main disadvantage is that the size of the feature vector is large. Accordingly, we use a configuration from autoregressive features (only the shaded pixels from the input window in Fig. 3), instead of all the pixels in the input window. This reduces the size of the feature vector from M^2 to $(4M-3)$, thereby resulting in an improved generalization performance and classification speed.

After training, the neural network outputs the real value between 0 and 1 for eye and non-eye respectively. If a pixel has a larger value than the given threshold value, it is considered as an eye; otherwise it is labeled as non-eye. Accordingly, the result of the classification is a binary image.

Fig. 4 shows the classification result using the neural network. Fig. 4(a) is an input frame and Fig. 4(b) is the extracted facial regions, and then the result of detected eye pixels is shown in Fig. 4(c). In Fig. 4(c), the pixels to be classified as eyes are marked as black. We can see that all of the eyes are labeled correctly, but there are some misclassified regions as eye.

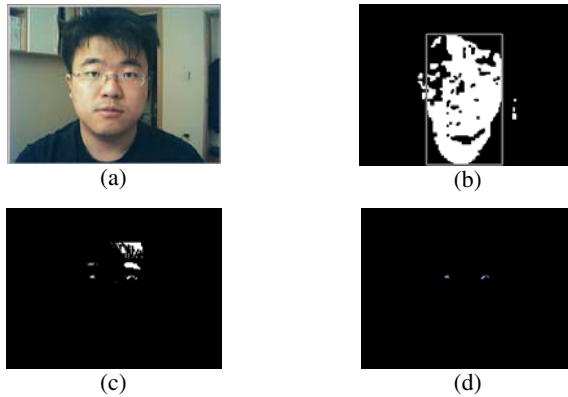


Fig. 4. An example of eye detection. (a) an original image, (b) the extracted facial regions (c) the classified image by the neural network, (d) the detected eye region after post-processing.

3.2 Post-processing

Although we use the bootstrap method to make the eye detection, the detection result from the MLP includes many false alarms. As such we still encounter difficulties in filtering out high-frequency and high-contrast non-eye regions. In this paper, we use the connected-component analysis result posterior to the texture classification. The generated connected-components (CC) are filtered by their attributes, such as size, area, and location. We have then applied two-stage filtering on the CCs:

Stage 1: Heuristics using features of CCs such as area, fill factor, and horizontal and vertical extents; The width of the eye region must be larger than the predefined minimum width (Min_width), the area of the component should be larger than Min_area and smaller than Max_area, and the fill factor should be larger than Min_fillfactor.

Stage 2: Heuristics using the geometric alignment of eye components; We check the rotation angles of two nearby eye candidates. They have to be smaller than the predefined rotation angle.

Using these two heuristics, the classified images are filtered, then the resulting image is shown in Fig. 4(d). In Fig. 4(d), the extracted eye region is filled blue for the better viewing.

4 Mean-Shift Based Eye Tracker

To track the detected eyes, a mean shift algorithm is used, which finds the object by seeking the mode of the object score distribution. In the present work, the color distribution of detected pupil, $P_m(g_s) = -(2\pi\sigma)^{-1/2} \exp\{-(g_s - \mu)^2 / \sigma^2\}$, where the μ and σ are set to 40 and 4 respectively. The distribution is used as the object score distribution at site s , which represents the probability of belonging to an eye.

A mean shift algorithm is a nonparametric technique that climbs the gradient of a probability distribution to find the nearest dominant mode. The algorithm iteratively shifts the center of the search window to the weighted mean until the difference between the means of successive iterations is less than a threshold. The depth of the color means the object score, i.e. the probability of belonging to an object. The solid-line rectangle is the search window for the current iteration, while the dotted line represents the shifted search window. The search window is moved in the direction of the object by shifting the center of the window to the weighted mean.

The weighted mean, i.e. the search window center at iteration $n+1$, m_{n+1} is computed using the following equation,

$$m_{n+1} = \sum_{s \in W} P_m(g_s) \cdot s / \sum_{s \in W} P_m(g_s) \tag{2}$$

The search window size for a mean shift algorithm is generally determined according to the object size, which is efficient when tracking an object with only a small motion. However, in many cases, objects have a large motion and low frame rate, which means the objects end up outside the search window. Therefore, a search window that is smaller than the object motion will fail to track the object. Accordingly, in this paper, the size of the search window of the mean shift algorithm is adaptively determined in direct proportion to the motion of the object as follows:

$$W_{width}^{(t)} = \max(\alpha(|m_x^{(t-1)} - m_x^{(t-2)}| - B_{width}), 0) + \beta B_{width} \tag{3}$$

$$W_{height}^{(t)} = \max(\alpha(|m_y^{(t-1)} - m_y^{(t-2)}| - B_{height}), 0) + \beta B_{height} \quad (t > 2)$$

where α and β are constant and t is the frame index. This adaptation of the window size allows for accurate tracking of highly active objects.

Fig. 5 shows the results of the eye tracking, where the eyes are filled out white for the better viewing. As can be seen in Fig. 5, the eye regions are accurately tracking. Moreover, the proposed method can determine the gaze direction.



Fig. 5. An eye tracking result

5 Experimental Results

To show the effectiveness of the proposed method, it was applied to the interface to convey users' command via an eye movement.

Fig. 6 shows the interface using our method, which consists of a PC camera and a computer. The PC camera, which is connected to the computer through the USB port,

supplies 30 color images of size 320×240 per second. The computer is a PentiumIV–1.7GHz with the Window XP operating system, and then it translates the user’s eye movements into the mouse movements by processing the images received from the PC camera. Then the processing of a video sequence is performed by the proposed eye tracking method.



Fig. 6. The interface system using our eye tracking method

To assess the effectiveness of the proposed method, the interface system was tested with 25-users under the various environments. The tracking results are shown in Fig. 7. The extracted eyes have been filled white for better viewing. The features are tracked throughout the 100 frames and not lost once.

To quantitatively evaluate the performance of the interface system, it was tested using an ‘alien game’ [11]. Each user was given an introduction to how the system worked and then allowed to practice moving the cursor for five minutes. The five-minute practice period was perceived as sufficient training time by the users. In the computer game, “aliens” appear at random locations on the screen one at a time, as shown in Fig. 8. To catch aliens, users must point the cursor at the boxes. Each user was asked to “catch ten aliens” three times with the standard mouse and three times with the mouse using eye movement. The type of mouse that was tested first was chosen randomly. For each test, the user’s time to play the game was recorded.

Table 1 presents the average time to be taken to catch one alien, when playing with the regular mouse and the mouse using eye movements. The former is about 3-times faster than the latter. But, the latter can process more than 30 frames/sec on a notebook without any additional hardware, for the 320×240 size input image, which is enough to apply to the real-time application. Moreover, the implemented system is not needed any additional hardware except a general PC and an inexpensive PC camera, the system is very efficient to realize many applications using real-time interactive information between users and computer systems.

Consequently, the experiment showed that it has a potential to be used as interface for handicapped people and generalized user interface in many applications.



Fig. 7. Tracking results in the cluttered environments

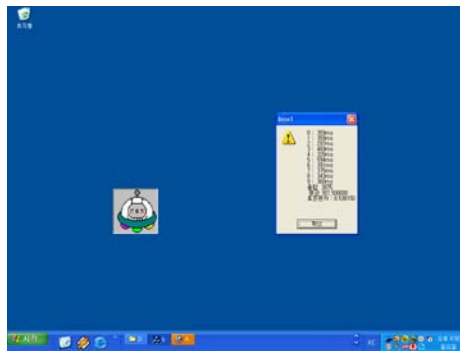


Fig. 8. Aliens game

Table 1. Timing comparison between regular mouse and the mouse using eye movements

Method	Measure	Time/sec
Standard Mouse	Mean	0.44s
	Deviation	0.07s
Eye Mouse	Mean	1.67s
	Deviation	0.21s

6 Conclusions

In this paper, we proposed an eye tracking method using NN and mean-shift procedure, and implemented the mouse system to receive user's eye as an input signal to control a computer. The proposed eye tracking method consists of three modules: face extractor, eye detector, and eye tracker. To deal with the rigid motion of a user, a face is first extracted using skin-color model and connected-component analysis. Thereafter, the user's eyes are localized by the NN-based texture classifier, and then the eyes are continuously tracking by a mean-shift procedure.

The proposed method was tested with 25 peoples, and the results shows that our method has the following advantages: 1) it is robust to the time-varying illumination and less sensitive to the specula reflection, 2) it works well on the input image of the low resolutions. However, the proposed method has some problems. Although it is fast enough to apply for user interface and other application, the proposed method is slower than the standard mouse. Therefore, we are currently investing the speed-up of our method.

References

1. Jacob, Robert J. K.: Human-computer interaction: Input devices. *ACM Computing Surveys*, Vol. 28, No. 1 (1996)
2. Sharma, R., Pavlovic, V.I., Huang, T.S.: Toward multimodal human-computer interface. *Proceedings of the IEEE*, Volume: 86, Issue: 5 (1998) 853 – 869
3. Kaufman, Arie E., Bandopadhyay, Amit., Shaviv, Bernard D.: An Eye Tracking Computer User Interface. *Virtual Reality, 1993. Proceedings., IEEE 1993 Symposium on Research Frontiers in*, 25-26 (1993)
4. Scassellati, Brian.: Eye finding via face detection for a foveated, active vision system. *American Association for Artificial Intelligence*. (1998)
5. Kurata, Takeshi., Okuma, Takashi., Kourogi, Masakatsu., Sakaue, Katsuhiko.: The Hand Mouse: GMM Hand-color Classification and Mean Shift Tracking. In *Proc. Second International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems (RATFG-RTS 2001) in conjunction with ICCV 2001 in Vancouver, Canada*. (2001) 119-124
6. Takami, N. Irie, Kang C., Ishimatsu, T., Ochiai, T.: Computer interface to use head movement for handicapped people. *TENCON '96. Proceedings. 1996 IEEE TENCON. Digital Signal Processing Applications*, Volume: 1, 26-29 Nov. vol. 1 (1996) 468 – 472

7. Sako, H., Whitehouse, M., Smith, A., Sutherland, A.: Real-time facial-feature tracking based on matching techniques and its applications. Pattern Recognition, 1994. Vol. 2 - Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on , Volume: 2 , 9-13 vol.2 (1994) 320 - 324
8. Kim, Sang-Hoon., Kim, Hyoung-Gon., Tchah, Kyun-Hyon.: Object oriented face detection using range and color information. Electronics Letters , Volume: 34 , Issue: 10 , 14 (1998) 979 – 980
9. Schiele, Bernet., Waibel, Alex.: Gaze Tracking Based on Face-Color. School of Computer Science, Carnegie Mello University (1995)
10. Yang, Jie., A., Waibel.: A real-time face tracker. Applications of Computer Vision, 1996. WACV '96., Proceedings 3rd IEEE Workshop on , 2-4 (1996) 142 – 147
11. Betke, M., Gips, J., Fleming, P.: The camera mouse: visual tracking of body features to provide computer access for people with severe disabilities. Neural Systems and Rehabilitation Engineering, IEEE Transactions on [see also IEEE Trans. on Rehabilitation Engineering] , Volume: 10 , Issue: 1. (2002) 1 – 10

The Optimal Feature Extraction Procedure for Statistical Pattern Recognition

Marek Kurzynski and Edward Puchala

Wroclaw University of Technology, Faculty of Electronics, Chair of Systems and Computer Networks, Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
marek.kurzynski@pwr.wroc.pl, edward.puchala@pwr.wroc.pl

Abstract. The paper deals with the extraction of features for object recognition. Bayes' probability of correct classification was adopted as the extraction criterion. The problem with full probabilistic information is discussed in detail. A simple calculation example is given and solved. One of the paper's chapters is devoted to a case when the available information is contained in the so-called learning sequence (the case of recognition with learning).

1 Introduction

One of the fundamental problems in statistical pattern recognition is representing patterns through a reduced number of dimensions. In most practical cases, the pattern feature space dimension is rather large due to the fact that it is too difficult or impossible to directly evaluate the usefulness of a particular feature at the design stage. Therefore it is reasonable to initially include all the potentially useful features and to reduce this set later.

There are two main methods of dimensionality reduction ([1], [2], [6]): *feature selection* in which we select the best possible subset of input features and *feature extraction* consisting in finding a transformation to a lower dimensional space. We shall concentrate here on feature reduction.

There are many effective methods of feature reduction. One can consider here linear and nonlinear feature extraction methods, particularly ones which ([4], [5]):

1. minimize the conditional risk or probability of incorrect object classification,
2. maximize or minimize the previously adopted objective function,
3. maximize the criteria for the information values of the individual features (or sets of features) describing the objects.

In each of the above cases, extraction means a feature space transformation leading to a reduction in the dimensionality of the space. One should note that feature extraction may result in deterioration of object classification quality and so it should be performed taking into account an increase in computer system operating speed and a reduction in data processing time while maintaining the best possible quality of the decision aiding systems.

In this paper a novel approach to the statistical problem of feature reduction is proposed. A linear transformation of the n -dimensional space of features

was adopted as the base for the extraction algorithm. The space allows one to describe an object in a new m -dimensional space with reduced dimensionality ($m < n$). In order to define a linear transformation one should determine the values of the transformation matrix components. This means that a properly defined optimisation problem should be solved. The probability of correct classification of objects in the recognition process was adopted as the criterion.

2 Preliminaries and the Problem Statement

Let us consider the pattern recognition problem with probabilistic model. This means that n -dimensional vector of features describing recognized pattern $x = (x_1, x_2, \dots, x_n)^T \in \mathcal{X} \subseteq \mathcal{R}^n$ and its class number $j \in \mathcal{M} = \{1, 2, \dots, M\}$ are observed values of a couple of random variables (\mathbf{X}, \mathbf{J}) , respectively. Its probability distribution is given by *a priori* probabilities of classes

$$p_j = P(\mathbf{J} = j), \quad j \in \mathcal{M} \quad (1)$$

and class-conditional probability density function (CPDFs) of \mathbf{X}

$$f_j(x) = f(x/j), \quad x \in \mathcal{X}, \quad j \in \mathcal{M}. \quad (2)$$

In order to reduce dimensionality of feature space let consider linear transformation

$$y = Ax, \quad (3)$$

which maps n -dimensional input feature space \mathcal{X} into m -dimensional derivative feature space $\mathcal{Y} \subseteq \mathcal{R}^m$, or - under assumption that $m < n$ - reduces dimensionality of space of object descriptors. It is obvious, that y is a vector of observed values of m dimensional random variable \mathbf{Y} , which probability distribution given by CPDFs depends on mapping matrix A , viz.

$$g(y/j; A) = g_j(y; A), \quad y \in \mathcal{Y}, \quad j \in \mathcal{M}. \quad (4)$$

Let introduce now a criterion function $Q(A)$ which evaluates discriminative ability of features y , i.e. Q states a measure of feature extraction mapping (3). As a criterion Q any measure can be involved which evaluates both the relevance of features based on a feature capacity to discriminate between classes or quality of a recognition algorithm used later to built the final classifier. In the further numerical example the Bayes probability of correct classification will be used, namely

$$Q(A) = Pc(A) = \int_{\mathcal{Y}} \max_{j \in \mathcal{M}} \{p_j g_j(y; A)\} dy. \quad (5)$$

Without any loss of generality, let us consider a higher value of Q to indicate a better feature vector y . Then the feature extraction problem can be formulated as follows: for given *priors* (1), CPDFs (2) and reduced dimension m find the matrix A^* for which

$$Q(A^*) = \max_A Q(A). \quad (6)$$

3 Optimization Procedure

In order to solve (6) first we must explicitly determine CPDFs (4). Let introduce the vector $\bar{y} = (y, x_1, x_2, \dots, x_{n-m})^T$ and linear transformation

$$\bar{y} = \bar{A} x, \tag{9}$$

where

$$\bar{A} = \begin{bmatrix} A & & \\ - & - & - \\ I & | & 0 \end{bmatrix} \tag{8}$$

is a square matrix $n \times n$. For given y equation (7) has an unique solution given by Cramer formulas

$$x_k(y) = |\bar{A}_k(y)| \cdot |\bar{A}|^{-1}, \tag{9}$$

where $\bar{A}_k(y)$ denotes matrix with k -th column replaced with vector \bar{y} . Hence putting (9) into (2) and (4) we get CPDFs of \bar{y} ([3]):

$$\bar{g}_j(\bar{y}; A) = J^{-1} \cdot f_j(x_1(\bar{y}), x_2(\bar{y}), \dots, x_n(\bar{y})), \tag{10}$$

where J is a Jacobian of mapping (7). Integrating (10) over variables x_1, \dots, x_{n-m} we simply get

$$g_j(y; A) = \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \dots \int_{\mathcal{X}_{n-m}} \bar{g}_j(\bar{y}; A) dx_1 dx_2 \dots dx_{n-m}. \tag{11}$$

Formula (11) allows one to determine class-conditional density functions for the vector of features y , describing the object in a new m -dimensional space. Substituting (11) into (5) one gets a criterion defining the probability of correct classification for the objects in space \mathcal{Y} :

$$\begin{aligned} Q(A) = Pc(A) &= \int_{\mathcal{Y}} \max_{j \in \mathcal{M}} \left\{ p_j \cdot \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \dots \int_{\mathcal{X}_{n-m}} J^{-1} \times \right. \\ &\quad \left. \times f_j(x_1(\bar{y}), x_2(\bar{y}), \dots, x_n(\bar{y})) dx_1 dx_2 \dots dx_{n-m} \right\} dy = \\ &= \int_{\mathcal{Y}} \max_{j \in \mathcal{M}} \left\{ p_j \cdot \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \dots \int_{\mathcal{X}_{n-m}} J^{-1} \times \right. \\ &\quad \left. \times f_j(|\bar{A}_1(y)| \cdot |\bar{A}|^{-1}, \dots, |\bar{A}_n(y)| \cdot |\bar{A}|^{-1}) dx_1 dx_2 \dots dx_{n-m} \right\} dy. \tag{12} \end{aligned}$$

Thus, the solution of the feature extraction problem (6) requires that such matrix A^* should be determined for which the Bayes probability of correct classification (12) is the maximum one.

Consequently, complex multiple integration and inversion operations must be performed on the multidimensional matrices in order to obtain optimal values of A . Although an analytical solution is possible (for low n and m values), it is complicated and time-consuming. Therefore it is proposed to use numerical procedures. For linear problem optimisation (which is the case here) classic numerical algorithms are very ineffective. In a search for a global extremum they have to be started (from different starting points) many times whereby the time needed to obtain an optimal solution is very long. Thus it is only natural to use the parallel processing methodology offered by genetic algorithms ([7]). A major difficulty here is the proper encoding of the problem. Chapter 4 presents a simple example of optimisation associated with the extraction of features, solved using the analytical method.

4 Numerical Example

Let consider two-class pattern recognition task with equal *priors* and reduction problem of feature space dimension from $n = 2$ to $m = 1$. Input feature vector is uniformly distributed and its CPDFs are as follows:

$$f_1(x) = \begin{cases} 0.5 & \text{for } 0 \leq x_1 \leq 2 \text{ and } 0 \leq x_2 \leq x_1, \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

$$f_2(x) = \begin{cases} 0.5 & \text{for } 0 \leq x_1 \leq 2 \text{ and } x_1 \leq x_2 \leq 2, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Now feature extraction mapping (3) has now the form

$$y = [a, 1] \cdot [x_1, x_2]^T = a \cdot x_1 + x_2 \tag{15}$$

and problem is to find such a value a^* which maximize criterion (12).

Since Jacobian of (7) is equal to 1 hence from (9) and (10) for $j = 1, 2$ we get

$$\bar{g}_j(\bar{y}, a) = f_j(x_1, y - a \cdot x_1). \tag{16}$$

The results of integrating (16) over x_1 , i.e. CPDFs (11) for $a \geq 1, -1 \leq a \leq 1$ and $a \leq -1$ are presented in Fig.1. a), b) and c), respectively.

Finally, from (5) we easy get:

$$P_c(a) = \begin{cases} \frac{a+1}{4a} & \text{for } a \geq |1|, \\ \frac{a+1}{4} & \text{for } a \leq |1|. \end{cases} \tag{17}$$

The graph demonstrating the Bayes probability of misclassification $P_e(a) = 1 - P_c(a)$ depending on parameter a of feature extraction mapping is depicted in Fig.2. The best result $P_e(a^*) = 0$ (or equivalently $P_c(a^*) = 1$) is obtained for $a^* = -1$.

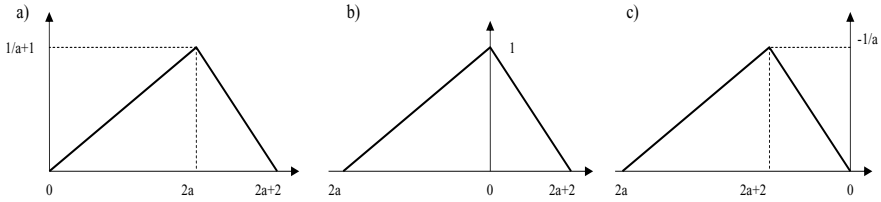


Fig. 1. Illustration of example

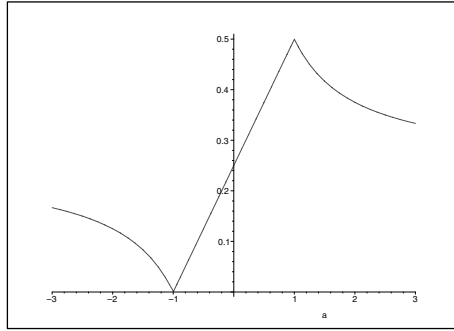


Fig. 2. Probability of misclassification

5 The Case of Recognition with Learning

It follows from the above considerations that an analytical and numerical solution of the optimisation problem is possible. But for this one must know the class-conditional density functions and the *a priori* probabilities of the classes. In practice, such information is rarely available. All we know about the classification problem is usually contained in the so-called learning sequence:

$$S_L(x) = \{(x^{(1)}, j^{(1)}), (x^{(2)}, j^{(2)}), \dots, (x^{(L)}, j^{(L)})\}. \tag{18}$$

Formula (18) describes objects in space \mathcal{X} . For the transformation to space \mathcal{Y} one should use the relation:

$$y^{(k)} = A \cdot x^{(k)}; \quad k = 1, 2, \dots, L \tag{19}$$

and then the learning sequence assumes the form:

$$S_L(y) = \{(y^{(1)}, j^{(1)}), (y^{(2)}, j^{(2)}), \dots, (y^{(L)}, j^{(L)})\}. \tag{20}$$

The elements of sequence $S_L(y)$ allow one to determine (in a standard way) the estimators of the *a priori* probabilities of classes p_{jL} and class-conditional density functions $f_{jL}(x)$. Then the optimisation criterion assumes this form:

$$Q_L(A) = P_{c_L}(A) = \int_{\mathcal{Y}} \max_{j \in \mathcal{M}} \left\{ p_{jL} \cdot \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \dots \int_{\mathcal{X}_{n-m}} J^{-1} \times \right. \\ \left. \times f_{jL}(x_1(\bar{y}), x_2(\bar{y}), \dots, x_n(\bar{y})) dx_1 dx_2 \dots dx_{n-m} \right\} dy. \quad (21)$$

6 Conclusions

The feature extraction problem is fundamental for recognition and classification tasks since it leads to a reduction in the dimensionality of the feature space in which an object is described. This is highly important in situations when it is essential that the computer system should aid on-line decision taking. In most cases, an optimisation task must be formulated for the extraction problem. Because of the high computational complexity involved, numerical methods are used for this purpose. The authors propose to apply genetic algorithms in the considered case. Simulation studies justifying this approach and their results will be the subject of subsequent publications.

References

1. Devroye L., Györfi P., Lugosi G.: A Probabilistic Theory of Pattern Recognition, Springer Verlag, New York, 1996
2. Duda R., Hart P., Stork D.: Pattern Classification, Wiley-Interscience, New York, 2001
3. Golub G., Van Loan C.: Matrix Computations, Johns Hopkins University Press, 1996
4. Guyon I., Gunn S., Nikravesh M., Zadeh L.: Feature Extraction, Foundations and Applications, Springer Verlag, 2004
5. Park H., Park C., Pardalos P.: Comparative Study of Linear and Nonlinear Feature Extraction Methods - Technical Report, Minneapolis, 2004
6. Fukunaga K.: Introduction to Statistical Pattern Recognition, Academic Press, 1990.
7. Goldberg D.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, New York, 1989

A New Approach for Human Identification Using Gait Recognition

Murat Ekinçi

Computer Vision Lab, Department of Computer Engineering,
Karadeniz Technical University, Trabzon, Turkey
ekinçi@ktu.edu.tr

Abstract. Recognition of a person from gait is a biometric of increasing interest. This paper presents a new approach on silhouette representation to extract gait patterns for human recognition. Silhouette shape of a motion object is first represented by four 1-D signals which are the basic image features called the distance vectors. The distance vectors are differences between the bounding box and silhouette. Second, eigenspace transformation based on Principal Component Analysis is applied to time-varying distance vectors and the statistical distance based supervised pattern classification is then performed in the lower-dimensional eigenspace for recognition. A fusion task is finally executed to produce final decision. Experimental results on three databases show that the proposed method is an effective and efficient gait representation for human identification, and the proposed approach achieves highly competitive performance with respect to the published gait recognition approaches.

1 Introduction

Human identification from gait has been a recent focus in computer vision. It is a behavioral biometric source that can be acquired at a distance. Gait recognition aims to discriminate individuals by the way they walk and has the advantage of being non-invasive, hard to conceal, being readily captured without a walker's attention, and is less likely to be obscured than other biometric features [1][2][3][6].

Gait recognition can be broadly divided into two groups, model-based and silhouette-based methods. Model-based methods [2][15] model the human body structure and extract image features to map them into structural components of models or to derive motion trajectories of body parts. The silhouette-based methods [6][7][9][1], characterizes body movement by the statistics of the patterns produced by walking. These patterns capture both the static and dynamic properties of body shape.

In this paper an effective representation of silhouette for gait recognition is developed and statistical analysis is performed. Similar observations have been made in [7][9][1], but the idea presented here implicitly captures both structural (appearances) and transitional (dynamics) characteristics of gait. The silhouette-based method presented is basically to produce the distance vectors, which are four 1D signals extracted from projections to silhouette, they are top-, bottom-,

left-, and right-projections. As following main purpose, depending on four distance vectors, PCA based gait recognition algorithm is first performed. A statistical distance based similarity is then achieved to obtain similarity measures on training and testing data. A fusion task includes two strategies is executed to produce consolidation decision. Experimental results on three different databases demonstrate that the proposed algorithm has an encouraging recognition performance.

2 Silhouette Representation

To extract spatial silhouettes of walking figures, a background modeling algorithm and a simple correspondence procedure are first used to segment and track the moving silhouettes of a walking figure, more details are given in [5]. Once a silhouette generated, a bounding box is placed around silhouette. Silhouette across a motion sequence (a gait cycle) are automatically aligned by scaling and cropping based on the bounding box. The details on the gait cycle estimation used are given in reference [14].

Silhouette representation is based on the projections to silhouette which is generated from a sequence of binary silhouette images $bs(t) = bs(x, y, t)$, indexed spatially by pixel location (x, y) and temporally by time t . There are four different image features called the distance vectors. They are top-, bottom-, left- and right-distance vectors. The distance vectors are the differences between the bounding box and the outer contour of silhouette. An example silhouette and the distance vectors corresponding to four projections are shown in the middle of figure 1. The distance vectors are separately represented by four 1D signals. The size of 1D signals is equal to the height or to the width of the bounding box for left- and right-distance vectors or for top- and bottom-distance vectors, respectively. The values in the signals for both left- and right-projections are computed as the difference in the locations of the bounding box and left-most and right-most boundary pixels, respectively, in a given row. The other projections along a given column are also computed as the differences from the top of the bounding box to the top-most of silhouette for top-projection, from the bottom of the box to the bottom-most of silhouette pixels for bottom-projections, respectively.

From a new 2D image $F^T(x, t) = \sum_y bs(x, y, t)$, where each column (indexed by time t) is the top-projections (row sum) of silhouette image $bs(t)$, as shown in figure 1 top-left. Each value $F^T(x, t)$ is then a count of the number of the row pixels between the top side of the bounding box and the outer contours in that columns x of silhouette image $bs(t)$. The result is a 2D pattern, formed by stacking top projections together to form a spatio-temporal pattern. A second pattern which represents the bottom-projection $F^B(x, t) = \sum_{-y} bs(x, y, t)$ can be constructed by stacking bottom projections, as shown in figure 1 bottom-left. The third pattern $F^L(y, t) = \sum_x bs(x, y, t)$ is then constructed by stacking with using the differences as column pixels from left side of the box to left-most boundary pixels of silhouette which are produced by the left projections, and the last pattern $F^R(y, t) = \sum_{-x} bs(x, y, t)$ is also finally constructed by stacking the right projections, as shown in figure 1 top-right and bottom-right 2D patterns,

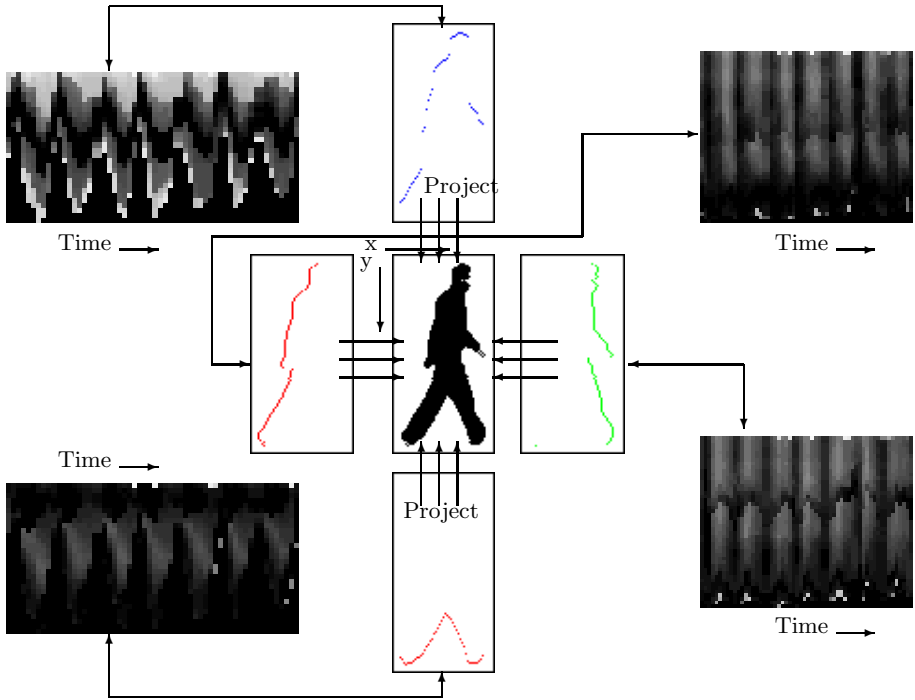


Fig. 1. Silhouette representation. **(Middle)** Silhouette and four projections, **(Left)** temporal plot of the distance vectors for top and bottom projections, **(Right)** temporal plot of the distance vectors for left and right projections.

respectively. The variation of each component of the each distance vectors can be regarded as gait signature of that object. From the temporal distance vector plots, it is clear that the distance vector is roughly periodic and gives the extent of movement of different part of the subject. The brighter a pixel in 2D patterns in figure 1, the larger value is the value of the distance vector in that position.

3 Training

The following processes on the four 1D signals produced from the distance vectors are to eliminate the influence of spatial scale and signal length of the distance vectors by scaling of these distance vector signals with respect to magnitude and size through the sizes of the bounding boxes. Eigenspace transformation based on Principal Component Analysis (PCA) is then applied to time varying distance vectors derived from a sequence of silhouette images to reduce the dimensionality of the input feature space. The training process similar to [1][4] is illustrated as follows:

Given k class for training, and each class represents a sequence of the distance vector signals of a person. Multiple sequences of each subject can be added for

training, but a sequence includes one gait cycle was considered in the experiments. Let $V_{i,j}^w$ be the j th distance vector signal in the i th class for w projection to silhouette and N_i the number of such distance vector signals in the i th class. The total number of training samples is $N_t^w = N_1^w + N_2^w + \dots + N_k^w$, as the whole training set can be represented by $[V_{1,1}^w, V_{1,2}^w, \dots, V_{1,N_1}^w, V_{2,1}^w, \dots, V_{k,N_k}^w]$. The mean m_v^w and the global covariance matrix \sum^w of w projection training set can easily be obtained by

$$m_v^w = \frac{1}{N_t^w} \sum_{i=1}^k \sum_{j=1}^{N_i^w} V_{i,j}^w \quad (1)$$

$$\sum^w = \frac{1}{N_t^w} \sum_{i=1}^k \sum_{j=1}^{N_i^w} (V_{i,j}^w - m_v^w)(V_{i,j}^w - m_v^w)^T \quad (2)$$

Here each V^w value is 1D signal and equal to, $F^w(\cdot)$, the distance vectors for w projection (top-bottom-left-right) as explained in section 2. If the rank of matrix \sum is N , then the N nonzero eigenvalues of \sum , $\lambda_1, \lambda_2, \dots, \lambda_N$, and associated eigenvectors e_1, e_2, \dots, e_N can be computed based on theory of *singular value decomposition* [4]. The first few eigenvectors correspond to large changes in training patterns, and higher-order eigenvectors represent smaller changes [1]. As a result, for computing efficiency in practical applications, those small eigenvalues and their corresponding eigenvectors are ignored. Then a transform matrix $T^w = [e_1^w, e_2^w, \dots, e_s^w]$ to project an original distance vector signal $V_{i,j}^w$ into a point $P_{i,j}^w$ in the eigenspace is constructed by taking only $s < N$ largest eigenvalues and their associated eigenvectors for each projections to silhouette. Therefore, s values are usually much smaller than the original data dimension N . Then the projection average A_i^w of each training sequence in the eigenspace is calculated by averaging of $P_{i,j}^w$ as follows:

$$P_{i,j}^w = [e_1^w \ e_2^w \ \dots \ e_s^w]^T V_{i,j}^w, \quad A_i^w = \frac{1}{N_i} \sum_{j=1}^{N_i} P_{i,j}^w \quad (3)$$

4 Pattern Classification

Gait pattern recognition (classification) can be solved through measuring similarities between reference gait pattern and test samples in the parametric eigenspace. A simple statistical distance has been chosen to measure similarity, because the main interest here is to evaluate the genuine discriminatory ability of the extracted features in the proposed method. The accumulated distance between the associated centroids A^w (obtained in the process of training) and B^w (obtained in the process of testing) can be easily computed by

$$d_S(A, B) = \sqrt{\left(\frac{A_1 - B_1}{s_1}\right)^2 + \dots + \left(\frac{A_p - B_p}{s_p}\right)^2} \quad (4)$$

Where (s_1, \dots, s_p) are equal to corresponding the sizes of A_i and B_i . In the distance measure, the classification result for each projection is then accomplished by choosing the minimum of d . The classification process is carried out via the nearest neighbor (NN) classifier. The classification is performed by classifying in the test sequence and all training sequences by

$$c = \arg_i \min d_i(B, A_i) \quad (5)$$

where B represents a test sequence, A_i represents the i th training sequence, d is the similarity measures described in above.

The similarity results produced from each distance vectors are fused to increase the recognition performance. In this fusion task, two different strategies were developed. In the **strategy 1**, each projection is separately treated. Then the strategy is combining the distances of each projections at the end by assigning equal weight. As implementation, if any two of the similarities achieved based on four projections give maximum similarities for same individual, then the identification is appointed as positive. This fusion strategy has rapidly increased the recognition performance in the experiments.

At the experiments, it has been seen that, some projection has given more robust results than others. For instance, while human moves in lateral view with respect to image plane, the back side of human can give more individual characteristics in gait. So, the projection corresponding to that side can give more reliable results. It is called dominant feature to this case. As second strategy, **the strategy 2** has also been developed to further increase the recognition performance. In the strategy 2, if the projection selected as dominant feature or at least two projections of others give positive for an individual, then identification result given by the strategy 2 is appointed as positive. The dominant feature in this work is automatically assigned by estimating the direction of motion objects in tracking. At the next section, the dominant features determined by experimentally for different view points with respect to image plane are given.

5 Experimental Results

The performance of the proposed methods was evaluated on CMU's MoBo database[13], NLPR gait database [1], and USF database [6]. The Viterbi algorithm was used to identify the test sequence, since it is efficient and can operate in the logarithmic domain using only additions [12]. The performance of the algorithm is evaluated on three different databases of varying of difficulty.

CMU Database. This database has 25 subjects (23 males, 2 females) walking on a treadmill. Each subject is recorded performing four different types of walking: slow walk, fast walk, inclined walk, and slow walk holding ball. There are about 8 cycles in each sequence, and each sequences is recorded at 30 frames per second. It also contains six simultaneous motion sequence of 25 subjects, as shown in figure 2.

One of the cycle in each sequence was used for testing, others for training. First, we did the following experiments on this database: **1)** train on slow walk

Table 1. Classification performance on the CMU data set for viewpoint 1

Test/Train	All projections: equal				Dominant: Right projection		
	Rank 1	Rank 2	Rank 3	Rank 4	Rank 1	Rank 2	Rank 3
Slow/Slow	72	100	100	100	84	100	100
Fast/Fast	76	100	100	100	92	100	100
Ball/Ball	84	100	100	100	84	100	100
Slow/Fast	36	92	100	100	52	100	100
Fast/Slow	20	60	100	100	32	88	100
Slow/Ball	8	17	33	58	42	96	100
Fast/Ball	4	13	33	67	17	50	88
Ball/Slow	8	17	38	67	33	88	100
Ball/Fast	13	29	58	92	29	63	100

and test on slow walk, **2)** train on fast walk and test on fast walk, **3)** train on walk carrying a ball and test on walk carrying a ball, **4)** train on slow walk and test on fast walk, **5)** train on slow walk and test on walk carrying a ball, **6)** train on fast walk and test on slow walk, **7)** train on fast walk and test on walk carrying a ball, **8)** train on walk carrying a ball and test on slow walk, **9)** train on walk carrying a ball and test on fast walk.

The results obtained using the proposed method are summarized on the all cases **1)-9)** in Table 1. It can be seen that the right person in the top two matches 100% of times for the cases where testing and training sets correspond to the same walk styles. When the strategy developed in the fusion as dominant feature (projections) is used, the recognition performance is increased, as seen in Table 1. For the case of training with fast walk and testing on slow walk, and vice versa, the dip in performance is caused due to the fact that for some individual as biometrics suggests, there is a considerable change in body dynamics and stride length as a person changes his speed. Nevertheless, the right person in the top three matches 100% of the times for that cases, and dominant projection strategy has also increased the recognition performance for Ranks 1 and 2. For the case of training with walk carrying ball and testing on slow and fast walks, and vice versa, encouraging results have also been produced by using the proposed method, and the dominant feature property has still increased the recognition performance, as given in Table 1.

**Fig. 2.** The six CMU database viewpoints

Table 2. Classification performance on the CMU data set for all views. Eight gait cycles were used, seven cycles for training, one cycle for testing.

View	Test/Train	All projections equal			Dominant: Right projection		
		Rank 1	Rank 2	Rank 3	Rank 1	Rank 2	Rank 3
4	Slow/Slow	76	100	100	84	100	100
	Fast/Fast	84	100	100	96	100	100
	Slow/Fast	12	44	80	24	64	100
	Fast/Slow	20	64	100	32	76	100
					Dominant: Left projection		
5	Slow/Slow	80	100	100	80	100	100
	Fast/Fast	88	100	100	88	100	100
	Slow/Fast	16	44	80	24	64	100
	Fast/Slow	24	56	96	32	68	100
					Dominant: Right projection		
3	Slow/Slow	80	100	100	88	100	100
	Fast/Fast	72	100	100	76	100	100
	Slow/Fast	20	64	100	28	76	100
	Fast/Slow	24	56	92	28	68	100
					Dominant: Right projection		
6	Slow/Slow	72	100	100	84	100	100
	Fast/Fast	76	100	100	80	100	100
	Slow/Fast	16	44	88	36	76	100
	Fast/Slow	16	40	72	24	56	100

For the other view points, the experimental results are also summarized on the cases **1)-4)** in Table 2. When the all experimental results for the different view points are considered, it can be seen that, the right person in the top two matches 100% and in the top four matches 100% of the times for the cases **1)-2)** and for the cases **3)-4)**, respectively. It is also seen that, when the dominant feature is used, gait recognition performance is also increased. Some comparison results are also given in Table 3. The reason to show the cases **1)-6)** only is become the page limitation in this paper, and that points are our lowest results on MoBo dataset for the comparisons to the other works in literature.

NLPR Database. The *NLPR* database [1] includes 20 subjects and four sequences for each viewing angle per subject, two sequences for one direction of

Table 3. Comparison of several algorithm on MoBo dataset

Algorithms	train/test	Rank 1(%)	Rank 2(%)	Rank 3 (%)	Rank 5 (%)
The method presented	slow/slow	84	100	100	100
Kale <i>et.al.</i> [7]	slow/slow	72	80	85	97
Collins <i>et.al.</i> [11]	slow/slow	86	100	100	100
The method presented	fast/slow	52	100	100	100
Kale <i>et.al.</i> [7]	fast/slow	56	62	75	82
Collins <i>et.al.</i> [11]	fast/slow	76	Not	given	92

Table 4. Performance on the NLPR data set for three views

Walking Direction	View	Training	Test	Rank1	Rank2	Rank3
One Way Walking	Lateral	Exp. 1	Exp. 1	65	100	100
		Exp. 1	Exp. 2	55	100	100
	Frontal	Exp. 1	Exp. 1	60	100	100
		Exp. 1	Exp. 2	35	100	100
	Oblique	Exp. 1	Exp. 1	40	90	100
		Exp. 1	Exp. 2	30	60	100
Reverse Way Walking	Lateral	Exp. 1	Exp. 1	60	100	100
		Exp. 1	Exp. 2	50	100	100
	Frontal	Exp. 1	Exp. 1	60	100	100
		Exp. 1	Exp. 2	40	100	100
	Oblique	Exp. 1	Exp. 1	45	100	100
		Exp. 1	Exp. 2	35	75	100

walking, the other two sequences for reverse direction of walking. For instance, when the subject is walking laterally to the camera, the direction of walking is from right to left for two of four sequences, and from right to left for the remaining. Those all gait sequences were captured as twice (we called two experiments) on two different days in an outdoor environment. All subjects walk along a straight-line path at free cadences in three different views with respect to the image plane, as shown in figure 3, where the white line with arrow represents one direction path, the other walking path is reverse direction.

We did the following experiments on this database: **1)** train on one image sequence and test on the remainder, all sequences were produced from first experiment, **2)** train on two sequences obtained from first experiment and test on two sequences obtained from second experiment. This is repeated for each viewing angle, and for each direction of walking. The results for the experiments along with cumulative match scores in three viewing angle are summarized in Table 4. When the experimental results are considered, the right person in the top two matches 100% of times for lateral and frontal viewing angles, and in the top three matches 100% of times for oblique view.

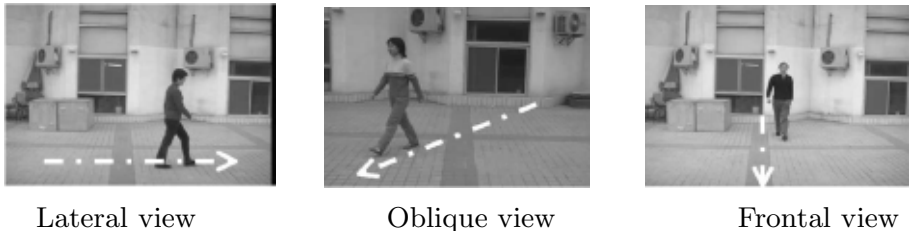


Fig. 3. Some images in the NLPR database

In the experiments on the NLPR database, the performance of the proposed algorithm was also compared with those of a few recent silhouette-based methods described in [11],[8],[16], and [1], respectively. To some extent, they reflect the latest and best work of these research groups in gait recognition. In [11], a method based on template matching of body silhouettes in key frames for human identification was established. The study in [8] described a moment-based representation of gait appearance for the purpose of person identification. A baseline algorithm was also proposed for human identification using spatio temporal correlation of silhouette images in [16]. The work in [1] computes the centroid of silhouette’s shape, and unwraps the outer counter to obtain a 1D distance signal, then applies principal component analysis for person identification. These methods were implemented using the same silhouette data from the NLPR database with lateral view by the study in [1], and the results given are as taken from tables in [1]. Table 5 lists the identification rates that have been reported by other algorithms and our algorithm. The proposed algorithm has successfully given the right person in top two matches 100% the times for the NLPR database.

USF Database. Finally, the USF database [6] is considered. The database has variations as regards viewing direction, shoe type, and surface type. At the experiments, one of the cycle in each sequence was used for testing, others (3-4 cycles) for training. Different probe sequences for the experiments along with the cumulative match scores are given in Table 6 for the algorithm presented in this paper and three different algorithms [16][1][7]. The same silhouette data from USF were directly used. These data are noisy, e.g., missing of body parts, small

Table 5. Comparison of Several algorithm on the NLPR Database (Lateral View)

Methods	Rank 1(%)	Rank 2(%)	Rank 3 (%)	Rank 5 (%)	Rank 10 (%)
BenAbdelkader [10]	73	Not given	89	96	
Collins [11]	71	Not given	79	88	
Lee [8]	88	Not given	99	100	
Phillips [16]	79	Not given	91	99	
Wang [1]	75	Not given	98	100	
The methods presented	65	100	100	100	100

Table 6. Performance on the USF database for four algorithm

Exp.	The method			Baseline[16]		NLPR[1]		UMD[7]	
	Rank 1	Rank 2	Rank 3	Rank 1	Rank 5	Rank 1	Rank 5	Rank 1	Rank 5
GAL[68]	35	80	100	79	96	70	92	91	100
GBR[44]	34	82	100	66	81	58	82	76	81
GBL[44]	25	55	91	56	76	51	70	65	76
CAL[68]	39	90	100	30	46	27	38	24	46
CAR[68]	30	66	100	29	61	34	64	25	61
CBL[41]	30	78	100	10	33	14	26	15	33
CBR[41]	29	66	100	24	55	21	45	29	39
GAR[68]	34	60	90	-	-	-	-	-	-

holes inside the objects, severe shadow around feet, and missing and adding some parts around the border of silhouettes due to background characteristics. In Table 6, G and C indicate grass and concrete surfaces, A and B indicate shoe types, and L and R indicate left and right cameras, respectively. The number of subjects in each subset is also given in square bracket. It is observed that, the proposed method has given the right person in top three matches 100% of the times for training and testing sets corresponding to the same camera.

6 Conclusion

The method presented has given very close results to the existing works for Rank 1 on the databases tested, nevertheless it has almost given 100% accuracy for Ranks 2 and 3 on the all databases used. Nonlinear discriminant analysis will be developed as next study to achieve higher accuracy than the current for Rank 1.

References

1. L. Wang, T. Tan, H. Ning, W. Hu, Silhouette Analysis-Based Gait Recognition for Human Identification. IEEE Trans. on PAMI Vol.25, No. 12, Dec.,2003.
2. C. BenAbdelkader, R. G. Cutler, L. S. Davis, Gait Recognition Using Image Self-Similarity. EURASIP Journal of Applied Signal Processing, April, 2004.
3. G. V. Veres, *et. al*, What image information is important in silhouette-based gait recognition? Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2004.
4. P. Huang, C. Harris, M. Nixon, Human Gait Recognition in Canonical Space Using Temporal Templates. IEE Proc. Vision Image and Signal Proc. Conf., 1999.
5. M. Ekinici, E. Gedikli, Background Estimation Based People Detection and Tracking for Video Surveillance. Springer LNCS 2869, ISCIS 2003, November, 2003.
6. S. Sarkar, *et al* The HumanID Gait Challenge Problem: Data Sets, Performance, and Analysis. IEEE Trans. on Pat. Anal. and Mach. Intell., Vol.27, No. 2, 2005.
7. A. Kale, *et. al.*, Identification of Humans Using Gait. IEEE Trans. on Image Processing, Vol.13, No.9, September 2004.
8. L. Lee, W. Grimson, Gait Analysis for Recognition and Classification. Proc. IEEE, Int. Conference on Automatic Face and Gesture Recognition, pp. 155-162, 2002.
9. Yanxi Liu, R. T. Collins, T. Tsin, Gait Sequence Analysis using Frieze Patterns. Proc. of European Conf. on Computer Vision, 2002.
10. C. BenAbdelkader, *et.al*, Stride and Cadence as a Biometric in Automatic Person Identification and Verification. Proc. Int. Conf. Aut. Face and Gesture Recog.,2002.
11. R. Collins, R. Gross, and J. Shi, Silhouette-Based Human Identification from Body Shape and Gait. Proc. Int. Conf. Automatic Face and Gesture Recognition, 2002.
12. J. Phillips *et.al*, The FERET Evaluation Methodology for Face recognition Algorithm. IEEE Trans. Pattern Analysis and Machine Intell., vol.22, no.10, Oct.2000.
13. R. Gross, J. Shi, The CMU motion of body (MOBO) database. Tech. Rep. CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, June 2001.
14. M. Ekinici, E. Gedikli A Novel Approach on Silhouette Based Human Motion Analysis for Gait Recognition. ISVC 2005, LNCS 3804, pp.219-226, December 2005.
15. A. I. Bazin, M. S. Nixon, Gait Verification Using Probabilistic Methods. IEEE Workshop on Applications of Computer Vision, 2005.
16. P. Phillips, *et.al.*, Baseline Results for Challenge Problem of Human ID using Gait Analysis. Proc. Int. Conf. Automatic Face and Gesture Recognition, 2002.

Author Index

- Abbas, Cláudia Jacy Barenco V-819
Abraham, Ajith IV-40
Adamidis, Panagiotis V-108
Adolf, David I-711
Ahn, Byung Jun II-77
Ahn, Dong-In III-251
Ahn, Heejune IV-603
Ahn, Jaehoon V-522
Ahn, Jung-Chul IV-370
Ahn, ManKi III-48
Ahn, Sang-Ho III-279
Ahn, Seongjin II-400, II-410, II-487,
V-829, II-1169
Ahn, Sung-Jin II-982
Ahn, Sungsoo V-269
Ahn, Sungwoo II-175
Ahn, Taewook IV-388
Ahn, Yonghak V-1001
Ahn, Youngjin II-661
Akbar, Ali Hammad II-186, II-847
Alam, Muhammad Mahbub II-651
Albertí, Margarita I-721
Alfredo-Badillo, Ignacio III-456
Ali, Hassan IV-217
Ali, Saqib IV-217
Allayear, Shaikh Muhammad II-641
Almendra, Daniel V-819
Al-Mutawah, Khalid I-586
Alvarez, Susana III-1073
Amarnadh, Narayanasetty I-1
Ambler, Anthony P. V-531
An, Kyoungwan II-155
An, Sunshin II-730
Anagun, A. Sermet III-11, III-678
Anan, Yoshiyuki II-40
Andersen, Anders Magnus IV-98
Angelides, Marios C. IV-118
Arce-Santana, Edgar R. V-412
Armer, Andrey I-974
Arteconi, Leonardo I-694

Badea, Bogdan I-1166
Bae, Hae-Young I-914, IV-1126
Bae, Hyo-Jung I-151
Bae, Joonsoo II-379
Bae, Suk-Tae II-309
Bae, Yong-Geun IV-828
Bae, Youngchul III-244
Baek, Jun-Geol V-839
Baek, Myung-Sun V-752
Bagherpour, Morteza III-546
Bahn, Hyokyung I-1072
Bai, Zhang I-885
Baik, Heung Ki V-236
Baixauli, J. Samuel III-1073
Bala, Piotr V-394
Balas, Lale I-547
Balci, Birim I-373
Ban, Chaehoon II-175
Bang, Hyungbin II-319
Bang, Young-Cheol III-1090, III-1129
Bardhan, Debabrata I-10
Bartolotta, Antonino I-821
Bashir, Ali Kashif II-186
Basu, Kalyan I-566
Bawa, Rajesh Kumar I-1177
Bellaachia, Abdelghani V-346
Bentz, Cédric III-738
Berbegall, Vicente V-192
Berkov, Dmitri V-129
Bertoni, Guido III-1004
Biscarri, Félix V-725
Biscarri, Jesús V-725
Blibech, Kaouthar III-395
Boada, Imma I-364
Bohli, Jens-Matthias III-355
Bolze, Raphael V-202
Bories, Benoît I-744
Bravo, Maricela IV-169
Brennan, John K. V-743
Breveglieri, Luca III-1004
Brzeziński, Jerzy IV-1166, V-98
Buiati, Fabio V-819
Burns, John I-612
Byun, Doyoung V-537
Byun, Sang-Seon II-1189
Byun, Sang-Yong III-84
Byun, Sung-Wook I-232

- Byun, Tae-Young III-134
 Byun, Young Hwan V-457
 Byun, Yung-Cheol V-185
 Byun, Yung-Hwan V-512, V-932

 Caballero-Gil, Pino I-577, III-1035
 Caballero, Ismael III-984
 Cáceres, Santos II-18
 Cai, Guoyin IV-1090
 Calderon, Alejandro IV-1136
 Calero, Coral III-984
 Camahort, Emilio I-510
 Camara, José Sierra V-798
 Campos-Delgado, Daniel U. V-412
 Cao, Wenming V-375
 Capacho, Liliana III-554
 Cappelletti, David I-721
 Carballeira, Félix García V-108
 Carlone, Pierpaolo I-794
 Caro, Angelica III-984
 Caron, Eddy V-202
 Carretero, Jesus IV-1136
 Castro, Mildrey Carbonell V-798
 Cattani, Carlo I-785, I-828, I-857
 Cha, Byung-Rae II-1090
 Cha, Eui-Young I-1110
 Cha, Guang-Ho I-344
 Cha, Jae-Sang V-312
 Cha, JeongHee V-432
 Chae, Kijoon I-1072, IV-440
 Chae, Oksam V-1001
 Chae, Young Seok II-760
 Challiol, Cecilia IV-148
 Chan, Yung-Kuan V-384
 Chan, Yuen-Yan I-383, III-309, III-365,
 III-507, IV-406
 Chang, Chung-Hsien I-171
 Chang, Hangbae IV-255, IV-707
 Chang, Hoon IV-577
 Chang, Hoon V-1010
 Chang, Hsi-Cheng V-158
 Chang, Kuo-Hwa III-944
 Chang, Ok-Bae III-188, III-222, IV-893,
 IV-955, V-644
 Chang, Soo Ho II-451
 Chang, Sujeong II-77
 Chang, Yu-Hern III-649
 Chaudhry, Shafique Ahmad II-847
 Chaudhuri, Chitrita II-1
 Chen, Chiou-Nan IV-1107
 Chen, Chun V-39
 Chen, Gencai V-39
 Chen, Huifen III-944
 Chen, Kai-Hung V-384
 Chen, Kaiyun IV-756
 Chen, Ken I-307
 Chen, Lei II-1149
 Chen, Ling V-39
 Chen, Tzu-Yi III-1081
 Chen, Yen Hung III-631
 Cheng, Jingde III-1
 Cheng, Yu-Ming I-171, I-181
 Cheon, Saeng Hoon III-718
 Cheon, SeongKwon III-73
 Cheun, Du Wan II-451
 Cheung, Yen I-586
 Chi, Sang Hoon IV-58
 Chih, Wen-Hai III-668
 Chlebiej, Michał V-394
 Cho, Cheol-Hyung I-101
 Cho, Daerae IV-787
 Cho, Dongyoung IV-491
 Cho, Eun Sook II-1003, IV-985
 Cho, Haengrae V-214
 Cho, Ik-hwan I-326
 Cho, Jae-Hyun I-1110
 Cho, Jong-Rae III-832, III-994
 Cho, Juphil V-236
 Cho, KumWon V-522
 Cho, Kwang Moon IV-1003
 Cho, Mi-Gyung I-904
 Cho, Minju II-760
 Cho, Nam-deok V-546
 Cho, Sang-Hun II-288
 Cho, Sok-Pal II-1082
 Cho, Sung-eon V-600
 Cho, Tae Ho IV-58
 Cho, Yongyun IV-30
 Cho, Yookun II-701, IV-499, IV-549
 Cho, Youngsong I-111
 Cho, You-Ze II-631
 Choi, Bong-Joon V-912
 Choi, Byung-Cheon III-785
 Choi, Byungdo IV-808
 Choi, Byung-Sun II-945
 Choi, Chang IV-567
 Choi, Changyeol II-562, IV-1156
 Choi, Deokjai IV-128
 Choi, Eun Young IV-316
 Choi, Ho-Jin II-796

- Choi, Hwangkyu II-562, IV-1156
 Choi, Hyung-Il V-441
 Choi, Hyun-Seon III-728
 Choi, Jaeyoung IV-11, IV-30
 Choi, Jonghyoun II-525, II-895
 Choi, Jongmyung II-1033
 Choi, Jong-Ryeol IV-893
 Choi, Junho IV-567
 Choi, Junkyun V-829
 Choi, Kuiwon I-335
 Choi, Kyung Cheol IV-659
 Choi, Misook II-49, IV-966
 Choi, Sang-soo V-618
 Choi, Sang-Yule V-312, V-322, V-355
 Choi, Seongman III-222, IV-955, V-644,
 V-675
 Choi, Su-il II-77
 Choi, Sung-Hee IV-937
 Choi, Tae-Young I-307, I-317
 Choi, Wan-Kyoo IV-828
 Choi, Wonjoon IV-279
 Choi, Yeon Sung I-993
 Choi, Yong-Rak IV-432
 Choi, Yun Jeong II-298
 Chon, Jaechoon I-261, III-1172
 Chon, Sungmi II-28
 Choo, Hyunseung II-165, II-288, II-534,
 II-661, II-710, II-856, II-923, II-934,
 II-1121, III-1090, III-1129
 Choo, MoonWon IV-787
 Chun, Junchul I-410
 Chung, Chin Hyun I-929, I-964
 Chung, Chun-Jen III-862
 Chung, Ha Joong IV-549
 Chung, Hyoung-Seog V-491
 Chung, Il-Yong IV-828
 Chung, Jinwook II-487, II-982
 Chung, Kyoil III-375, IV-584, V-251
 Chung, Min Young II-77, II-288, II-534,
 II-856, II-934, II-1121
 Chung, Mokdong IV-1042
 Chung, Shu-Hsing III-610
 Chung, TaeChoong II-390
 Chung, Tae-sun I-1019
 Chung, Tai-Myoung II-135, II-239,
 III-486, V-626, V-655
 Chung, YoonJung III-54, IV-777
 Chung, Younky III-198, III-234
 Ciancio, Armando I-828
 Ciancio, Vincenzo I-821
 Clifford, Raphaël III-1137
 Cocho, Pedro III-964
 Cokuslu, Deniz II-681
 Coll, Narcis I-81
 Cong, Jin I-921
 Cools, Ronald V-780
 Cordero, Rogelio Limón IV-726
 Costantini, Alessandro I-738
 Crane, Martin I-612
 Cruz, Laura II-18
 Culley, Steve J. II-279
 Cumplido, René III-456
 Czekster, Ricardo M. I-202
 Daefler, Simon I-566
 Dagdeviren, Orhan II-681
 D'Anjou, Alicia III-1143
 Darlington, Mansur J. II-279
 Das, Gautam K. II-750
 Das, Sajal I-566
 Das, Sandip I-10, II-750
 David, Gabriel IV-78
 De Cristófolo, Valeria IV-148
 de Deus, Flavio E. V-808
 de Doncker, Elise V-789
 de Oliveira, Robson V-819
 de Frutos-Escrig, David IV-158
 de Ipiña, Diego López IV-108
 de Sousa, Rafael V-819
 Deineko, Vladimir III-793
 Demirkol, Askin V-365
 den Hertog, Dick III-812
 Deo, Puspita I-622
 Derevyankin, Valery I-974
 Desprez, Frederic V-202
 Dévai, Frank I-131
 Diaz, Olivia Graciela Fragoso IV-50
 Doallo, Ramón I-701
 Dogdu, Erdogan IV-88
 Drummond, L.A. V-192
 Duan, Guolin V-450
 Duan, Yucong IV-746
 Durán, Alfonso III-964
 Ekinci, Murat III-1216
 Eksioğlu, Burak III-748
 Eksioğlu, Sandra Duni III-708
 Eom, Jung-Ho II-239
 Eom, Young Ik I-1028
 Erciyes, Kayhan II-681

- Esquivel, Manuel L. III-841
 Eun, He-Jue V-990
 Evangelisti, Stefano I-744

 Fang, Zhijun II-964
 Färber, Gerrit III-638
 Farsaci, Francesco I-821
 Farzanyar, Zahra I-1100
 Fathy, Mahmood V-118
 Fei, Chai IV-179
 Feixas, Miquel I-449
 Feng, Dan I-1045
 Feregrino-Uribe, Claudia III-456
 Ferey, Nicolas I-222
 Fernandez, Javier IV-1136
 Fernández, Marcel III-527
 Fernández, Marcos I-490
 Fernandez, Reinaldo Togores I-30
 Fernández-Medina, Eduardo III-1013,
 III-1024, III-1044
 Fey, Dietmar V-129
 Filomia, Federico I-731
 Fiore, Ugo III-537
 Fleissner, Sebastian I-383, IV-406
 Forné, Jordi IV-1098
 Fort, Marta I-81
 Frausto, Juan IV-169
 Frick, Alexander I-847
 Fu, Xiaolan IV-746
 Fúster-Sabater, Amparo I-577, III-1035

 Gabillon, Alban III-395
 Galindo, David III-318
 Gallego, Guillermo III-822
 Gao, Yunjun V-39
 Garcia, Felix IV-1136
 Garcia, Jose Daniel IV-1136
 García, L.M. Sánchez V-108
 García-Sebastian, M. Teresa III-1143
 Gattiker, James R. III-1153
 Gattton, Thomas M. III-244, IV-947,
 V-665, V-675
 Gaudiot, Jean-Luc IV-622
 Gavrilova, Marina L. I-61, I-431
 Ge, He III-327
 Gerardo, Bobby D. III-144, IV-899,
 V-867
 Gervasi, Osvaldo I-212, I-665
 Gherbi, Rachid I-222
 Ghosh, Preetam I-566
 Ghosh, Samik I-566
 Gil, Joon-Min II-1169
 Go, Sung-Hyun V-867
 Goh, John I-1090
 Goh, Sunbok II-204
 Goi, Bok-Min IV-424
 González, Ana I. III-1143
 Gonzalez, César Otero I-30
 González, J.J. V-772
 González, Juan G. II-18
 González, Luis I-633
 González, Patricia I-701
 Gordillo, Silvia IV-148
 Górriz, J.M. V-772
 Graña, Manuel III-1143
 Gros, Pierre Emmanuel I-222
 Gu, Boncheol IV-499, IV-549
 Gu, Huaxi V-149
 Gu, Yuqing IV-746
 Guillen, Mario II-18
 Guo, Jiang III-974
 Guo, Jianping IV-1090
 Guo, Weiliang I-938
 Gutiérrez, Miguel III-964

 Ha, Jong-Eun III-1163
 Ha, Jongsung II-49
 Ha, Sung Ho III-1110
 Hahn, GeneBeck II-769
 Hamid, Md.Abdul II-866
 Han, Chang-Hyo III-832
 Han, DoHyung IV-594
 Han, Dong-Guk III-375
 Han, Gun Heui V-331
 Han, Hyuksoo IV-1081
 Han, Jizhong I-1010
 Han, Jong-Wook IV-360
 Han, Joohyun IV-30
 Han, JungHyun I-1028
 Han, Jungkyu IV-549
 Han, Ki-Joon II-259
 Han, Kijun II-1159
 Han, Kunhee V-584
 Han, Long-zhe I-1019
 Han, Sang Yong IV-40
 Han, Seakjae V-682
 Han, SeungJae II-359
 Han, Sunyoung II-601
 Han, Youngshin V-260
 Harbusch, Klaus I-857

- Hashemi, Sattar I-1100
 Hawes, Cathy I-644
 He, S. III-934
 Helal, Wissam I-744
 Heng, Swee-Huay III-416
 Heo, Joon II-989, II-1066
 Heo, Junyoung II-701, IV-499, IV-549
 Hérissou, Joan I-222
 Herrero, José R. V-762
 Higdon, David III-1153
 Hinarejos, M. Francisca IV-1098
 Hoesch, Georg V-202
 Hoffmann, Aswin L. III-812
 Hong, Bonghee II-155, II-175
 Hong, Choong Seon II-651, II-866
 Hong, Dong-Suk II-259
 Hong, Dowon IV-584
 Hong, Gye Hang III-1110
 Hong, Jiman II-701, IV-499, IV-558, IV-603
 Hong, John-Hee III-832
 Hong, Kwang-Seok I-354
 Hong, Maria II-400
 Hong, In-Hwa IV-245
 Hong, Seokhie III-446
 Hong, Soonjwa III-385
 Hong, Suk-Kyo II-847
 Hong, Sukwon I-1019
 Hong, Sung-Je I-151
 Hong, Sung-Pil III-785
 Hong, WonGi IV-577
 Hong, Youn-Sik II-249
 Horie, Daisuke III-1
 Hosaka, Ryosuke I-596
 Hsieh, Shu-Ming V-422
 Hsu, Chiun-Chieh V-158, V-422
 Hsu, Li-Fu V-422
 Hu, Qingwu IV-746
 Hu, Yincui IV-1090
 Huang, Changqin V-243
 Huang, Chun-Ying III-610
 Huang, Wei I-518
 Huh, Euinam II 390, II-515, II-827, II-905, V-717
 Huh, Woong II-224
 Huh, Woonghee Tim III-822
 Hur, Tai-Sung II-224
 Hwang, An Kyu II-788
 Hwang, Chong-Sun II-369, II-816
 Hwang, Ha-Jin V-1018
 Hwang, Hyun-Suk III-115, III-125, V-895
 Hwang, InYong II-1140
 Hwang, Jin-Bum IV-360
 Hwang, Jun II-760
 Hwang, Ken III-668
 Hwang, Soyeon IV-344
 Hwang, Suk-Hyung IV-767, IV-937
 Hwang, Sungho III-134
 Hwang, Sun-Myung IV-909
 Hwang, Tae Jin V-236
 Ikeguchi, Tohru I-596
 Im, Chaeseok I-1000
 Im, Eul Gyu III-54, IV-777
 Im, SeokJin II-369
 Im, Sungbin II-806
 Inan, Asu I-547
 Inceoglu, Mustafa Murat I-373
 Iordache, Dan I-804
 Isaac, Jesús Téllez V-798
 Isaila, Florin D. V-108
 Ishii, Naohiro II-40
 Iwata, Kazunori II-40
 Jang, Byung-Jun V-752
 Jang, Hyo-Jong II-106
 Jang, Injoo III-206
 Jang, Jun Yeong I-964
 Jang, Kil-Woong II-671
 Jang, Moonsuk II-204
 Jang, Sangdong IV-1116
 Jang, Taeuk II-760
 Jang, Yong-Il IV-1126
 Je, Sung-Kwan I-1110
 Jeon, Hoseong II-934
 Jeon, Hyung Joon II-974, II-1009
 Jeon, Hyung-Su III-188
 Jeon, Jongwoo III-718
 Jeon, Kwon-Su V-932
 Jeon, Segil V-522
 Jeon, Sung-Eok III-134
 Jeong, Byeong-Soo II-505, II-796
 Jeong, Chang-Sung I-232, II-462
 Jeong, Chang-Won IV-853
 Jeong, Chulho II-430
 Jeong, Dong-Hoon II-996
 Jeong, Dongseok I-326
 Jeong, Gu-Beom IV-1032
 Jeong, Hye-Jin V-675

- Jeong, Hyo Sook V-609
 Jeong, In-Jae III-698
 Jeong, Jong-Geun II-1090
 Jeong, Karpjoo V-522
 Jeong, Kugsang IV-128
 Jeong, KwangChul II-923
 Jeong, Sa-Kyun IV-893
 Jeong, Su-Hwan V-895
 Jeong, Taikyeong T. I-993, V-531
 Jhang, Seong Tae IV-631
 Jhon, Chu Shik IV-631
 Ji, JunFeng I-420
 Ji, Yong Gu IV-697
 Ji, Young Mu V-457
 Jiang, Chaojun I-938
 Jiang, Di I-50
 Jiang, Gangyi I-307, I-317
 Jiang, Yan I-921
 Jianping, Li I-885
 Jin, DongXue III-73
 Jin, Hai IV-529
 Jin, Honggee IV-687
 Jin, Mingzhou III-708, III-748
 Jo, Geun-Sik II-779
 Jo, Jeong Woo II-480
 Jodlbauer, Herbert V-88
 Johnstone, John K. I-500
 Joo, Su-Chong III-251, IV-798, IV-853,
 IV-899
 Joye, Marc III-338
 Juang, Wen-Shenq IV-396
 Ju, Hyunho V-522
 Ju, Minseong IV-271
 Jun, Jin V-839
 Jung, Cheol IV-687
 Jung, Eun-Sun IV-416
 Jung, Hyedong II-691
 Jung, Hye-Jung IV-1052
 Jung, Inbum II-562, IV-1156
 Jung, Jae-Yoon II-379, V-942
 Jung, JaeYoun III-64
 Jung, Jiwon II-155
 Jung, Won-Do II-186
 Jung, Kwang Hoon I-929
 Jung, Kyeong-Hoon IV-448
 Jung, Kyung-Hoon III-115
 Jung, Myoung Hee II-77
 Jung, SangJoon III-93, III-234, IV-1022
 Jung, Seung-Hwan II-462
 Jung, Se-Won II-837
 Jung, Won-Do II-186
 Jung, Won-Tae IV-1052
 Jung, Youngsuk IV-1022
 Jwa, JeongWoo IV-594
 Kabara, Joseph V-808
 Kangavari, Mohammadreza I-1100
 Kang, Dazhou II-1179
 Kang, Dong-Joong II-309, III-1163
 Kang, Dong-Wook IV-448
 Kang, Euisun II-400
 Kang, Euiyoung IV-558
 Kang, Eun-Kwan IV-947
 Kang, Heau-jo V-690
 Kang, Hong-Koo II-259
 Kang, Hyungwoo III-385
 Kang, Jeonil IV-380
 Kang, Jinsuk I-993
 Kang, Maing-Kyu III-898
 Kang, Mikyung IV-558
 Kang, Mingyun V-575
 Kang, Namhi III-497
 Kang, Oh-Hyung III-287, IV-1060
 Kang, Sanggil I-1127
 Kang, Sang-Won II-369
 Kang, Sangwook II-730
 Kang, Seo-Il IV-326
 Kang, Seoungpil II-1066
 Kang, Sin Kuk III-1200
 Kang, Suk-Ho IV-787, V-942
 Kang, Sukhoon II-1060, IV-271, IV-432
 Kang, Wanmo III-777, III-822
 Kang, Yunjeong V-665
 Karsak, E. Ertugrul III-918
 Kasprzak, Andrzej III-1100, III-1119
 Katsionis, George I-251
 Kaugars, Karlis V-789
 Keil, J. Mark I-121
 Képès, François I-222
 Kettner, Lutz I-60
 Key, Jaehong I-335
 Khader, Dalia III-298
 Khonsari, Ahmad V-118
 Khoo, Khoongming III-416
 Kim, Backhyun IV-68, IV-1146
 Kim, Bonghan V-851
 Kim, Bong-Je V-895
 Kim, Byeongchang III-21
 Kim, Byung Chul II-788

- Kim, Byunggi II-319, II-330, II-740,
 II-1033
 Kim, Byung-Guk II-996
 Kim, Byung-Ryong III-476
 Kim, Byung-Soon II-671
 Kim, Chang J. V-932
 Kim, Changmin III-261, IV-787
 Kim, Chang Ouk V-839
 Kim, Chang-Soo III-115, III-125, V-895
 Kim, Cheol Min I-278, I-288, IV-558
 Kim, Chonggun III-64, III-73, III-93,
 III-234, IV-808, IV-818, IV-1022
 Kim, Chulgoon V-522
 Kim, Chul Jin II-1003, IV-985
 Kim, Chul Soo V-185
 Kim, Dai-Youn III-38
 Kim, Deok-Soo I-101, I-111, I-440
 Kim, DongKook II-340, II-349
 Kim, Dong-Oh II-259
 Kim, Dong-Seok IV-853
 Kim, Dongsoo IV-687
 Kim, Donguk I-101, I-111, I-440
 Kim, Duckki II-195
 Kim, Duk Hun II-856
 Kim, Eung Soo III-31
 Kim, Eunhoe IV-11, IV-30
 Kim, Eun Mi III-1190, IV-893
 Kim, Eun Yi III-1200
 Kim, Gil-Han V-284
 Kim, Gui-Jung IV-835
 Kim, Guk-Boh IV-1032
 Kim, Gukboh II-214
 Kim, Gu Su I-1028
 Kim, Gwanghoon IV-344
 Kim, GyeYoung II-106, V-432, V-441
 Kim, Gyoung Bae I-914
 Kim, Hae Geun III-104
 Kim, Haeng-Kon III-84, III-163,
 III-198, IV-844, IV-927, IV-976
 Kim, Haeng Kon IV-873
 Kim, Hak-Jin III-928
 Kim, HanIl IV-558, IV-567, IV-594
 Kim, Hee Taek I-914
 Kim, Hong-Gee IV-937
 Kim, Hong-Jin II-1082
 Kim, Hong Sok V-1010
 Kim, Hong-Yeon I-1053
 Kim, Ho-Seok I-914, IV-1126
 Kim, Ho Won III-375
 Kim, Howon IV-584, V-251
 Kim, Hye-Jin I-955
 Kim, Hye Sun I-288
 Kim, HyoJin II-359
 Kim, Hyongsuk III-1172
 Kim, Hyun IV-466
 Kim, Hyuncheol V-829
 Kim, Hyung-Jun IV-483
 Kim, Hyunsoo III-852
 Kim, Iksoo IV-68, IV-1146
 Kim, IL V-912
 Kim, Ildo II-87
 Kim, InJung III-54, IV-777
 Kim, In Kee V-1
 Kim, Intae IV-21
 Kim, Jaehyoun II-934
 Kim, Jae-Soo II-572
 Kim, Jae-Yearn III-590
 Kim, Jaihie II-96
 Kim, Jee-In I-983
 Kim, Je-Min II-1219
 Kim, Jeong Hyun II-996, II-1066
 Kim, Jin-Geol IV-288
 Kim, Jin Ok I-929, I-964
 Kim, Jin Suk II-480
 Kim, Jin-Sung V-968, V-979
 Kim, Jin Won IV-499, IV-509
 Kim, John II-114
 Kim, Jong-Hwa V-503
 Kim, Jongik II-552
 Kim, Jongsung III-446
 Kim, Jongwan II-369
 Kim, June I-1028, I-1053
 Kim, Jungduk IV-255, IV-707
 Kim, Jung-Sun V-922
 Kim, Junguk II-760
 Kim, Kap Hwan III-564
 Kim, Kibom III-385
 Kim, Ki-Chang III-476
 Kim, Ki-Doo IV-448
 Kim, Ki-Hyung II-186, II-847
 Kim, Ki-Uk V-895
 Kim, Ki-Young IV-612
 Kim, Kuinam J. II-1025
 Kim, Kwang-Baek I-1110, III-172,
 III-279, V-887
 Kim, Kwangsoo IV-466
 Kim, Kwanjoong II-319, II-1033
 Kim, Kyujung II-28
 Kim, Kyung-Kyu IV-255
 Kim, Kyung Tae IV-519

- Kim, LaeYoung II-1131
 Kim, Min Chan IV-669
 Kim, Min-Ji V-932
 Kim, Minsoo III-154, IV-697, V-269,
 V-922
 Kim, Minsu III-134
 Kim, Min Sung III-31
 Kim, Misun II-420, III-154
 Kim, Miyoung II-885
 Kim, MoonHae V-522
 Kim, MoonJoon IV-577
 Kim, Moonseong II-710, III-1054,
 III-1090, III-1129, V-626
 Kim, Myeng-Ki IV-937
 Kim, Myoung-Joon I-1053
 Kim, Myoung-sub V-700
 Kim, Myung Keun I-914
 Kim, Nam-Gyun I-241
 Kim, Pankoo IV-567
 Kim, Sangbok II-515
 Kim, Sangho V-491
 Kim, Sang-II III-728
 Kim, Sangjin IV-388
 Kim, Sangki II-87
 Kim, Sangkuk II-11
 Kim, Sangkyun IV-639, IV-716
 Kim, Seki III-1054
 Kim, Seoksoo II-1060, IV-271, V-565,
 V-575, V-584, V-591, V-700
 Kim, Seok-Yoon IV-612
 Kim, Seong Baeg I-278, I-288, IV-558
 Kim, Seungjoo II-954, V-858
 Kim, Seung Man I-480
 Kim, Seung-Yong IV-612
 Kim, Sijung V-851
 Kim, SinKyu II-769
 Kim, Soo Dong II-451, IV-736
 Kim, Soo Hyung IV-128
 Kim, Soon-gohn V-690
 Kim, Soon-Ho III-172
 Kim, So-yeon V-618
 Kim, Su-Nam IV-448
 Kim, Sungchan I-459
 Kim, Sung Jin V-609
 Kim, Sung Jo IV-669
 Kim, Sung Ki II-876
 Kim, Sung-Ryul III-1137
 Kim, Sung-Shick III-928
 Kim, SungSoo I-904
 Kim, Sungsuk IV-567
 Kim, Tae-Kyung II-135
 Kim, Taeseok I-1062
 Kim, Tai-hoon V-700
 Kim, Ung Mo II-165, IV-456
 Kim, Ungmo I-1028, V-139
 Kim, Won II-106
 Kim, Woo-Jae II-720
 Kim, Wu Woan IV-1116
 Kim, Yang-Woo II-905
 Kim, Yeong-Deok IV-271
 Kim, Yong-Hwa V-968
 Kim, Yong-Min II-340, II-349
 Kim, Yongsik IV-687
 Kim, Yong-Sung V-958, V-968, V-979,
 V-990
 Kim, Yong-Yook I-241
 Kim, Yoon II-562, IV-1156
 Kim, Young Beom II-515, II-827
 Kim, Youngbong IV-226
 Kim, Youngchul II-319, II-1033
 Kim, Younghan III-497
 Kim, Younhyun II-611
 Kim, Young-Kyun I-1053
 Kim, Youngrag III-64
 Kim, Young Shin V-457
 Kim, Youngsoo II-545
 Kim, Yunkuk II-730
 Knauer, Christian I-20
 Ko, Eung Nam IV-475
 Ko, Hyuk Jin II-165
 Ko, Il Seok V-331, V-338
 Ko, Kwangsun I-1028
 Ko, Kyong-Cheol IV-1060
 Kobusińska, Anna IV-1166
 Koh, Byoung-Soo IV-236, IV-245
 Koh, Kern I-1062
 Koh, Yunji I-1062
 Kohout, Josef I-71
 Kolingerová, Ivana I-71
 Kong, Jung-Shik IV-288
 Koo, Jahwan II-487
 Kosowski, Adrian I-141, I-161
 Koszalka, Leszek V-58
 Koutsonikola, Vassiliki A. II-1229
 Kozhevnikov, Victor I-974
 Krasheninnikova, Natalia I-974
 Krasheninnikov, Victor I-974
 Kreveld, Marc van I-20
 Krusche, Peter V-165
 Ku, Chih-Wei II-1210

- Ku, Hyunchul II-827
 Kurzynski, Marek III-1210
 Kwak, Jae-min V-600
 Kwak, Jin II-954
 Kwak, Jong Min V-338
 Kwak, Jong Wook IV-631
 Kwak, Keun-Chang I-955
 Kwon, Dong-Hee II-720
 Kwon, Dong-Hyuck III-38
 Kwon, Gihwon IV-1081, V-905
 Kwon, Jang-Woo II-309, V-887
 Kwon, Jungkyu IV-1042
 Kwon, Oh-Cheon II-552
 Kwon, Oh-Heum IV-306
 Kwon, Seungwoo III-928
 Kwon, Soo-Tae III-767
 Kwon, Taekyoung II-769, II-915
 Kwon, Tae-Kyu I-241
 Kwon, Yoon-Jung V-503
- Laganà, Antonio I-212, I-665, I-675,
 I-694, I-721, I-738, I-757
 Lago, Noelia Faginas I-731
 Lai, Jun IV-179
 Lai, Kin Keung I-518
 Lan, Joung-Liang IV-1107
 Laskari, E.C. V-635
 Lawrence, Earl III-1153
 Lazar, Bogdan I-779
 Lazzareschi, Michael III-1081
 Le, D. Xuan IV-207
 Lee, Amy Hsin-I III-610
 Lee, Bo-Hee IV-288
 Lee, Bongkyu IV-549, V-185
 Lee, Byung-kwan III-38, III-172,
 III-261
 Lee, Byung-Wook I-946, II-495
 Lee, Chae-Woo II-837
 Lee, Changhee I-440
 Lee, Changhoon III-446
 Lee, Changjin V-537
 Lee, Chang-Mog IV-1012
 Lee, Chang-Woo IV-1060
 Lee, Chien-I II-1210
 Lee, Chilgee V-260
 Lee, Chulsoo IV-777
 Lee, Chulung III-928
 Lee, Chung-Sub IV-798
 Lee, Dan IV-994
 Lee, Deok-Gyu IV-326, IV-370
- Lee, Deokgyu IV-344
 Lee, Dong Chun II-1017, II-1051,
 II-1082
 Lee, Dong-Ho III-728
 Lee, Dong Hoon III-385, IV-316
 Lee, DongWoo IV-197, IV-491
 Lee, Dong-Young II-135, III-486, V-626,
 V-655
 Lee, SungYoung II-390
 Lee, Eun Ser IV-1070, V-546, V-555
 Lee, Eung Ju IV-187
 Lee, Eunseok II-430, II-621, V-49
 Lee, Gang-soo V-618
 Lee, Gary Geunbae III-21
 Lee, Geon-Yeob IV-853
 Lee, Geuk II-1060
 Lee, Gigan V-952
 Lee, Gueesang IV-128
 Lee, Gun Ho IV-659
 Lee, Hanku V-522
 Lee, Ha-Yong IV-767
 Lee, Hong Joo IV-639, IV-716
 Lee, HoonJae III-48, III-269
 Lee, Hosin IV-255
 Lee, Ho Woo III-718
 Lee, Hyewon K. II-214
 Lee, Hyobin II-96
 Lee, Hyun Chan I-111
 Lee, Hyung Su II-691, IV-519
 Lee, Hyung-Woo V-284, V-294
 Lee, Ig-hoon I-1036
 Lee, Im-Yeong IV-326, IV-370
 Lee, Inbok III-1137
 Lee, Jaedeuk V-675
 Lee, Jae Don I-1000
 Lee, Jae-Dong IV-1126
 Lee, Jae-Kwang II-945
 Lee, Jae-Seung II-945
 Lee, Jaewan III-144, III-178,
 IV-899, V-867
 Lee, Jae Woo V-457, V-512, V-932
 Lee, Jaewook II-487
 Lee, Jae Yeol IV-466
 Lee, Jaeyeon I-955
 Lee, Jae Yong II-788
 Lee, Jangho I-983
 Lee, Jang Hyun II-1199
 Lee, Jeong Hun III-600
 Lee, Jeonghyun IV-21
 Lee, JeongMin IV-577

- Lee, Ji-Hyun III-287, IV-994
 Lee, Jin Ho III-875
 Lee, Joahyoung II-562, IV-1156
 Lee, Jongchan II-1033
 Lee, Jong Gu III-1190
 Lee, Jong Sik V-1
 Lee, Jong-Sool III-564
 Lee, Jong-Sub III-898
 Lee, Jongsuk II-49
 Lee, Jungho I-326
 Lee, Junghoon IV-558, V-185
 Lee, Jungsuk V-269
 Lee, Junsoo V-175
 Lee, Kang-Hyuk II-309
 Lee, Kang-Woo IV-466
 Lee, Kang-Yoon II-827
 Lee, Keun-Ho II-816
 Lee, Keun Wang II-1074
 Lee, Kihyung II-175
 Lee, Kil-Hung II-572
 Lee, Kilsup IV-917, V-877
 Lee, Ki-Young II-249
 Lee, Kunwoo I-459
 Lee, Kwang Hyoung II-1074
 Lee, Kwangyong IV-499
 Lee, Kwan H. I-480
 Lee, Kwan-Hee I-151
 Lee, Kyesan II-905, V-708, V-717,
 V-952
 Lee, Kyujin V-708
 Lee, Kyu Min IV-483
 Lee, KyungHee IV-380
 Lee, Kyung Ho II-1199
 Lee, Kyunghye II-410
 Lee, Kyungsik III-777
 Lee, Kyung Whan IV-873
 Lee, Malrey III-244, IV-947, V-644,
 V-665, V-675
 Lee, Mun-Kyu IV-584
 Lee, Myungho I-1019
 Lee, Myungjin I-1072
 Lee, Na-Young V-441
 Lee, Samuel Sangkon II-231
 Lee, Sangjin IV-245
 Lee, Sang-goo I-1036
 Lee, Sang Ho IV-1070, V-555, V-609
 Lee, Sang-Hun II-239
 Lee, Sang Hun I-459
 Lee, Sangjin III-446, IV-236
 Lee, Sang Joon V-185
 Lee, Sang-Jun V-503
 Lee, Sangjun IV-549
 Lee, Sang-Min I-1053
 Lee, Sangyoung II-87, II-96
 Lee, Seojeong IV-966
 Lee, Seok-Cheol III-115
 Lee, Seon-Don II-720
 Lee, SeongHoon IV-491
 Lee, Seonghoon IV-197
 Lee, Seong-Won IV-622
 Lee, Seoung-Hyeon II-945
 Lee, Seoung-Soo V-503
 Lee, SeoungYoung II-1140
 Lee, Seungbae V-467
 Lee, Seung-Heon II-495
 Lee, Seunghwa II-621, V-49
 Lee, Seunghwan III-64, III-93
 Lee, Seung-Jin V-512
 Lee, Seungkeun IV-21
 Lee, Seungmin V-476
 Lee, Seung-Yeon II-905
 Lee, SeungYong V-922
 Lee, Se Won III-718
 Lee, SooCheol II-552
 Lee, SuKyoung II-1131
 Lee, Su Mi IV-316
 Lee, Sungchang II-923
 Lee, Sung-Hyup II-631
 Lee, Sung Jong IV-917
 Lee, Sung-Joo IV-828
 Lee, SungYoung II-390
 Lee, Sungkeun II-204
 Lee, Tae-Dong II-462
 Lee, Taehoon IV-1081
 Lee, Tae-Jin II-288, II-534, II-661,
 II-710, II-856, II-923, II-1121
 Lee, Vincent I-586
 Lee, Wankwon IV-491
 Lee, Wansuk II-954
 Lee, Wongoo II-11, V-851
 Lee, Won-Hyuk II-982
 Lee, Wookey IV-787, V-942
 Lee, Woongho I-326
 Lee, Yang-sun V-600
 Lee, Yongjin IV-197
 Lee, Yongseok V-665
 Lee, YoungGyo III-54
 Lee, Young Hoon III-875
 Lee, Young-Koo II-505
 Lee, Youngkwon II-915

- Lee, Young-Seok III-144
 Lee, Youngsook III-517, V-858
 Lee, YoungSoon V-675
 Lee, Yun Ho III-875
 Lee, Yun-Kyoung I-232
 Leininger, Thierry I-744
 León, Carlos V-725
 Leong, Chee-Seng IV-424
 Lho, Tae-Jung II-309, III-1163, V-887
 Liang, Yanchun I-938
 Liang, Zhong III-928
 Liao, Yuehong III-974
 Li, Fucui I-317
 Li, Haisen S. V-789
 Li, Jie II-59
 Li, Jin III-309, III-365, IV-406
 Li, Kuan-Ching IV-1107
 Li, Li I-895
 Li, Lv V-32
 Li, Qu I-393
 Li, Sheng I-420
 Li, Shiping I-317
 Li, Shujun V-789
 Li, Xun I-895
 Li, Yanhui II-1179
 Li, Yunsong V-149
 Li, Zhanwei V-450
 Li, Zhong I-1118, I-1134
 Lim, Andrew III-688
 Lim, Chan-Hyoung III-832
 Lim, Hyotaek IV-380
 Lim, Hyung-Jin II-135, II-239, III-486,
 V-626, V-655
 Lim, Jeong-Mi IV-679
 Lim, JiHyung IV-380
 Lim, Jiyoung IV-440
 Lim, Sungjun IV-707
 Lim, Taesoo IV-687
 Lim, Younghwan II-28, II-400,
 II-410, II-487
 Lin, Ching-Fen III-944
 Lin, Chuen-Horng V-384
 Lin, Hon-Ren V-158
 Lin, Hung-Mei III-338
 Lin, Kung-Kuei V-158
 Lin, Woei II-1111
 Ling, Yun IV-649
 Lísal, Martin V-743
 Lisowski, Dominik V-58
 Liu, Chia-Lung II-1111
 Liu, Fuyu IV-88
 Liu, Guoli III-659
 Liu, Heng I-528
 Liu, Joseph K. IV-406
 Liu, Jun IV-649
 Liu, Kai III-748
 Liu, Qun I-1045
 Liu, Shaofeng II-279
 Liu, Xianxing II-59
 Liu, XueHui I-420
 Loke, Seng Wai IV-138
 Lopes, Carla Teixeira IV-78
 López, Máximo IV-169
 Lu, Jiahui I-938
 Lu, Jianjiang II-1179
 Lu, Jiqiang III-466
 Lu, Xiaolin I-192, I-875
 Luna-Rivera, Jose M. V-412
 Luo, Ying IV-1090
 Luo, Yuan I-431
 Lv, Xinran V-450

 Ma, Hong III-688
 Ma, Lizhuang I-1118, I-1134
 Ma, Shichao I-1010
 Madern, Narcis I-81
 Magneau, Olivier I-222
 Mah, Pyeong Soo IV-509
 Makarov, Nikolay I-974
 Małafiejski, Michał I-141, I-161
 Mamun-or-Rashid, Md. II-651
 Manos, Konstantinos I-251
 Manzanares, Antonio Izquierdo V-798
 Mao, Zhihong I-1118, I-1134
 Markiewicz, Marta I-684
 Markowski, Marcin III-1119
 Marroquín-Alonso, Olga IV-158
 Martín, María J. I-701
 Mateo, Romeo Mark A. III-178, V-867
 Matte-Tailliez, Oriane I-222
 Maynau, Daniel I-744
 McLeish, Tom I-711
 McMahan, Chris A. II-279
 Mecke, Rüdiger I-268
 Meek, Dereck I-1118
 Mehlhorn, Kurt I-60
 Meletiou, G.C. V-635
 Mellado, Daniel III-1044
 Meng, Qingfan I-938
 Merabti, Madjid IV-352

- Mercorelli, Paolo I-847, I-857
 Miao, Zhaowei III-688
 Mijangos, Eugenio III-757
 Mikołajczak, Paweł V-394
 Millán, Rocío V-725
 Min, Byoung Joon II-270, II-876
 Min, Hong IV-499, IV-549
 Min, Hong-Ki II-224
 Min, Hyun Gi IV-736
 Min, Jun-Ki II-67
 Min, Kyongpil I-410
 Min, So Yeon II-1003, II-1074, IV-985
 Mitra, Pinaki I-1, II-1
 Moet, Esther I-20
 Moh, Chiou II-1111
 Mohades, Ali V-735
 Moin, M. Shahram III-1180
 Mok, Hak-Soo III-832, III-994
 Molinaro, Luis V-808
 Monedero, Íñigo V-725
 Moon, Aekyung V-214
 Moon, Il Kyeong III-600
 Moon, Ki-Young II-945
 Moon, Kwang-Sup III-994
 Moon, Mikyeong II-441, IV-226
 Moon, Young Shik V-404
 Morarescu, Cristian I-771, I-779,
 I-804, I-814, I-839
 Moreno, Anna M. Coves III-638
 Moreno, Ismael Solís IV-50
 Morillo, Pedro I-490
 Morimoto, Shoichi III-1
 Morphet, Steve I-1127
 Mouloudi, Abdelaaziz V-346
 Mouriño, Carlos J. I-701
 Mu, Yi III-345
 Mukhopadhyay, Sourav III-436
 Mukhtar, Shoab II-847
 Mun, Gil-Jong II-340
 Mun, YoungSong II-195, II-214, II-319,
 II-400, II-410, II-420, II-471, II-487,
 II-525, II-611, II-740, II-885, II-895
 Murzin, Mikhail Y. I-605
- Na, Yang V-467, V-476
 Na, Yun Ji V-331, V-338
 Nah, Jungchan IV-440
 Nakashima, Toyoshiro II-40
 Nam, Do-Hyun II-224
 Nam, Junghyun III-517, V-858
 Nam, Taekyong II-545
 Nandy, Subhas C. II-750
 Naumann, Uwe I-865
 Navarro, Juan J. V-762
 Neelov, Igor I-711
 Neumann, Laszlo I-449
 Ng, Victor I-383
 Niewiadomska-Szynkiewicz, Ewa I-537
 Nilforoushan, Zahra V-735
 Noh, Bong-Nam II-340, II-349, V-922
 Noh, Hye-Min III-188, IV-893
 Noh, Min-Ki II-1169
 Noh, Sang-Kyun II-349
 Noh, Seo-Young II-145
 Noh, SiChoon II-1051
 Noh, Sun-Kuk II-582
 Noori, Siamak III-546
 Noruzi, Mohammadreza III-1180
 Nowiński, Krzysztof V-394
 Nyang, DaeHun IV-380
- Ogryczak, Włodzimierz III-802
 Oh, Am-Suk II-309, V-887
 Oh, Hayoung IV-440
 Oh, Heekuck IV-388
 Oh, Hyukjun IV-603
 Oh, Inn Yeal II-974, II-1009
 Oh, Jaeduck II-471
 Oh, Jehwan V-49
 Oh, Juhyun II-760
 Oh, June II-1199
 Omary, Fouzia V-346
 Onosato, Masahiko I-469
 Orduña, Juan Manuel I-490
- Ortiz, Guillermo Rodríguez IV-50
 Ould-Khaoua, Mohamed V-118
- Paar, Christof III-1004
 Pacifici, Leonardo I-694
 Pahk, Cheryl Soo III-1190
 Paik, Juryon IV-456
 Pak, Jinsuk II-1159
 Palazzo, Gaetano Salvatore I-794
 Palmieri, Francesco III-537
 Pamula, Raj III-974
 Papadimitriou, Georgios I. II-1229
 Pardede, Eric I-1146, IV-207
 Park, Chang Mok IV-296

- Park, Chang-Seop IV-679
 Park, Chang Won IV-549
 Park, Chiwoo IV-697
 Park, Choung-Hwan II-1043
 Park, Dae-Hyeon V-958
 Park, DaeHyuck II-400
 Park, Dea-Woo IV-883
 Park, DongSik V-260
 Park, Eun-Ju IV-927
 Park, Geunyoung IV-549
 Park, Gilcheol V-565, V-591
 Park, Gi-Won II-631
 Park, Gyungleen II-760, IV-558, V-185
 Park, Hee-Un V-700
 Park, HongShik II-1140
 Park, Hyeong-Uk V-512
 Park, Ilgon IV-509
 Park, Jaehyung II-77
 Park, Jaekwan II-155
 Park, Jaemin IV-549
 Park, Jang-Su IV-370
 Park, Jea-Youn IV-835
 Park, Jeongmin II-430
 Park, Jeong Su IV-316
 Park, Jeung Chul I-480
 Park, Jong Hyuk IV-236, IV-245
 Park, Jongjin II-525
 Park, Joon Young I-111
 Park, Jungkeun I-1000
 Park, Jun Sang V-457
 Park, Ki-Hong IV-1060
 Park, Kisoeb III-1054
 Park, Ki Tae V-404
 Park, Kyoo-Seok III-125, V-912
 Park, Kyungdo III-928
 Park, Mee-Young V-512
 Park, Mi-Og IV-883
 Park, Namje V-251
 Park, Neungsoo IV-622
 Park, Sachoun V-905
 Park, Sangjoon II-319, II-1033
 Park, Sang Soon V-236
 Park, Sang Yong IV-1
 Park, SeongHoon V-68
 Park, Seungmin IV-499
 Park, Seung Soo II-298
 Park, Soo-Jin IV-432
 Park, Soon-Young IV-1126
 Park, Sung Soon II-641
 Park, Taehyung II-806
 Park, Wongil II-330
 Park, Woojin II-730
 Park, Yong-Seok IV-370
 Park, Young-Bae II-224
 Park, Young-Jae II-515
 Park, Young-Shin IV-432
 Park, Youngsup I-402
 Park, Young-Tack II-1219
 Parpas, Panos III-908
 Pastor, Rafael III-554
 Paun, Viorel I-779, I-804
 Pazo-Robles, Maria Eugenia I-577
 Pazos, Rodolfo II-18, IV-169
 Pegueroles, Josep III-527
 Pérez, Jesús Carretero V-108
 Pérez, Joaquín IV-169
 Pérez-Rosés, Hebert I-510
 Perrin, Dimitri I-612
 Petridou, Sophia G. II-1229
 Phillips, Robert III-822
 Piao, Xuefeng IV-549
 Piattini, Mario III-984, III-1013,
 III-1024, III-1044
 Pillards, Tim V-780
 Pineda-Rico, Ulises V-412
 Pion, Sylvain I-60
 Pirani, Fernando I-721, I-738
 Poch, Jordi I-364
 Pont, Michael J. V-22
 Pontvieux, Cyril V-202
 Porrini, Massimiliano I-721
 Porschen, Stefan I-40
 Pozniak-Koszalka, Iwona V-58
 Prados, Ferran I-364
 Puchala, Edward III-1210
 Pulcineli, L. V-819
 Puntonet, C.G. V-772
 Pusca, Stefan I-763, I-771, I-779,
 I-804, I-839
 Puttini, Ricardo Staciariini V-808
 Qin, Xujia I-393
 Qu, Xiangli V-224
 Qu, Zhiguo I-921
 Quintana, Arturo I-510
 Quirós, Ricardo I-510
 Rabenseifner, Rolf V-108
 Rahayu, J. Wenny I-1146, IV-207
 Rahman, Md. Mustafizur II-866

- Ramírez, J. V-772
 Rao, Imran II-390
 Reed, Chris I-644
 Rehfeld, Martina I-268
 Reitner, Sonja V-88
 Reyes, Gerardo IV-169
 Rhee, Choonsung II-601
 Rhee, Gue Won IV-466
 Rhee, Yang-Won III-287, IV-1060
 Rico-Novella, Francisco III-527
 Riganelli, Antonio I-665
 Rigau, Jaume I-449
 Rim, Kiwook IV-21
 Rodionov, Alexey S. I-605
 Roh, Byeong-hee IV-279
 Rosa-Velardo, Fernando IV-158
 Rossi, Gustavo IV-148
 Roy, Sasanka I-10
 Rubio, Monica I-510
 Ruskin, Heather J. I-612, I-622
 Rustem, Berç III-908
 Ryba, Przemyslaw III-1100
 Ryu, Jong Ho II-270
 Ryu, Yeonseung I-1000, I-1019
- Sadjadi, Seyed Jafar III-546, III-574
 Safaei, Farshad V-118
 Sakai, Yutaka I-596
 Salavert, Isidro Ramos IV-726
 Salgado, René Santaolaya IV-50
 Samavati, Faramarz I-91
 Sarac, T. III-678
 Sarkar, Palash III-436
 Saunders, J.R. I-556, III-934
 Sbert, Mateu I-449
 Schirra, Stefan I-60
 Schizas, Christos N. IV-118
 Schoor, Wolfram I-268
 Schurz, Frank V-129
 Ścisło, Piotr V-394
 Sedano, Iñigo IV-108
 Segura, J.C. V-772
 Sellarès, J. Antoni I-81
 Semé, David V-10
 Seo, Dae-Hee IV-326
 Seo, Dong Il II-270
 Seo, Dongmahn II-562, IV-1156
 Seo, JaeHyun III-154, V-922
 Seo, Jeongyeon I-101
- Seo, Kyu-Tae IV-288
 Seo, Manseung I-469
 Seo, Won Ju III-718
 Seo, Young-Jun IV-864
 Seo, Yuhwa III-954
 Severiano, José Andrés Díaz I-30
 Severn, Aaron I-91
 Shao, Feng I-307
 Shi, Qi IV-352
 Shi, Wenbo III-213
 Shibasaki, Ryosuke I-261
 Shim, Choon-Bo II-114
 Shim, Donghee IV-491
 Shim, Jang-Sup V-968, V-979, V-990
 Shim, Junho I-1036
 Shim, Young-Chul II-125, II-591
 Shimizu, Eihan I-261
 Shin, Chang-Sun III-251, IV-798
 Shin, Chungsoo II-740
 Shin, Dae-won III-261
 Shin, Dong-Ryeol IV-483
 Shin, Dong Ryul II-165
 Shin, Dongshin V-467
 Shin, Hayong I-440
 Shin, Ho-Jin IV-483
 Shin, Jeong-Hoon I-354
 Shin, Kee-Young IV-509
 Shin, Kwangcheol IV-40
 Shin, Myong-Chul V-312
 Shin, Seung-Jung II-487
 Shin, Woochul I-895
 Shin, Yongtae III-954
 Shuqing, Zhang I-885
 Siem, Alex Y.D. III-812
 Singh, David E. IV-1136
 Skouteris, Dimitris I-757
 Śliwiński, Tomasz III-802
 Smith, William R. V-743
 Sobaniec, Cezary V-98
 Sofokleous, Anastasis A. IV-118
 Soh, Ben IV-179
 Soh, Wooyoung V-682
 Sohn, Hong-Gyoo II-989, II-1043
 Sohn, Sungwon V-251
 Sohn, Surgwon II-779
 Soler, Emilio III-1024
 Soler, Josep I-364
 Son, Jeongho II-1159
 Son, Kyungho II-954
 Son, Seung-Hyun III-590

- Song, Eungkyu I-1062
 Song, Eun Jee II-1051
 Song, Ha-Joo IV-306
 Song, Hyoung-Kyu V-752
 Song, Jaekoo V-575
 Song, Jaesung I-469
 Song, JooSeok II-359, II-769, II-1131
 Song, Jungsuk IV-245
 Song, Ki Won IV-873
 Song, Sung Keun V-139
 Song, Wang-Cheol V-185
 Song, Yeong-Sun II-1043
 Song, Young-Jae IV-835, IV-864
 Soriano, Miguel III-527
 Sosa, Víctor J. II-18, IV-169
 Souza, Osmar Norberto de I-202
 Stanek, Martin III-426
 Stankova, Elena N. I-752
 Sterian, Andreea I-779, I-804
 Stewart, Neil F. I-50
 Storch, Lorian I-675
 Suh, Young-Ho IV-466
 Suh, Young-Joo II-720
 Sun, Jizhou V-450
 Sun, Lijuan V-450
 Sun, Youxian IV-539
 Sung, Jaechul III-446
 Sung, Sulyun III-954
 Susilo, Willy III-345
 Syukur, Evi IV-138
- Tae, Kang Soo II-231
 Takagi, Tsuyoshi III-375
 Talia, Domenico I-1080
 Tan, Pengliu IV-529
 Tan, Wuzheng I-1118, I-1134
 Tang, Chuan Yi III-631
 Tang, Lixin III-659
 Tang, W.J. I-556
 Taniar, David I-1090, I-1146
 Tarantelli, Francesco I-675
 Tasan, Seren Özmehmet V-78
 Tasoulis, D.K. V-635
 Tasso, Sergio I-212
 Tate, Stephen R. III-327
 Teng, Lirong I-938
 tie, Li V-32
 Tiskin, Alexander III-793, V-165
 Toma, Alexandru I-839
 Toma, Cristian I-779
 Toma, Ghiocel I-804
 Toma, Theodora I-771
 Torabi, Torab IV-98, IV-217
 Torres-Jiménez, José IV-726
 Tragha, Abderrahim V-346
 Trujillo, Juan III-1024
 Trunfio, Paolo I-1080
 Tsai, Cheng-Jung II-1210
 Tsai, Chwei-Shyong III-406
 Tsai, Yuan-Yu I-171, I-181
 Tunali, Semra V-78
- Uhm, Chul-Yong IV-448
- Vafadoost, Mansour III-1180
 Vakali, Athena I. II-1229
 Val, Cristina Manchado del I-30
 Vanhoucke, Mario III-621
 Varnuška, Michal I-71
 Vazquez, Juan Ignacio IV-108
 Vehreschild, Andre I-865
 Verdú, Gumersindo V-192
 Verta, Oreste I-1080
 Vidal, Vicente V-192
 Vidler, Peter J. V-22
 Villalba, Luis Javier García V-808,
 V-819
 Villarroel, Rodolfo III-1024
 Villarrubia, Carlos III-1013
 Villecco, Francesco I-857
 Virvou, Maria I-251
 Vlad, Adriana I-1166
 Voss, Heinrich I-684
 Vrahatis, M.N. V-635
- Walkowiak, Krzysztof II-1101
 Wan, Wei IV-1090
 Wan, Zheng II-964
 Wang, Bo-Hyun I-946
 Wang, Chung-Ming I-171, I-181
 Wang, Gi-Nam IV-296
 Wang, GuoPing I-420
 Wang, K.J. III-885
 Wang, Kun V-149
 Wang, Kung-Jeng III-668
 Wang, Shoujue V-375
 Wang, Shouyang I-518
 Wang, Weihong I-393
 Wang, Yanming III-309
 Wang, Zhengyou II-964

- Wang, Zhensong I-1010
 Watson, Mark D. I-121
 Wawrzyniak, Dariusz V-98
 Wee, Hui-Ming III-862, III-885
 Weidenhiller, Andreas V-88
 Wei, Guiyi IV-649
 Wei, Tao IV-262
 Wen, Chia-Hsien IV-1107
 Wheeler, Thomas J. I-654
 Wild, Peter J. II-279
 Wollinger, Thomas III-1004
 Won, Dong Ho II-165
 Won, Dongho II-545, II-954, III-54,
 III-517, IV-777, V-251, V-858
 Won, Youjip I-1062
 Woo, Sinam II-730
 Woo, Yo-Seop II-224
 Woo, Young-Ho II-224
 Wu, Chaolin IV-1090
 Wu, Chin-Chi II-1111
 Wu, EnHua I-420
 Wu, Hsien-Chu III-406
 Wu, Mary III-93, IV-818, IV-1022
 Wu, Q.H. I-556, III-934
 Wu, Qianhong III-345
 Wu, Shiqian II-964
 Wu, Xiaqing I-500
 Wu, Zhaohui II-1149
- Xia, Feng IV-539
 Xia, Yu III-1064
 Xiao, Zhenghong V-243
 Xiaohong, Li V-32
 Xie, Mei-fen V-375
 Xie, Qiming V-149
 Xie, Xiaoqin IV-756
 Xu, Baowen II-1179
 Xu, Fuyin V-243
 Xue, Yong IV-1090
- Yan, Jingqi I-528
 Yang, Byounghak III-581
 Yang, Ching-Wen IV-1107
 Yang, Hae-Sool IV-767, IV-937, IV-976,
 IV-1052
 Yang, Hwang-Kyu III-279
 Yang, Hyunho III-178
 Yang, Jong S. III-1129
 Yang, Kyoung Mi I-278
 Yang, Seung-hae III-38, III-261
- Yang, Xuejun V-224
 Yang, Young-Kyu II-495
 Yang, Young Soon II-1199
 Yap, Chee I-60
 Yeh, Chuan-Po III-406
 Yeh, Chung-Hsing III-649
 Yeo, So-Young V-752
 Yeom, Hee-Gyun IV-909
 Yeom, Keunhyuk II-441, IV-226
 Yeom, Soon-Ja V-958
 Yeun, Yun Seog II-1199
 Yi, Sangho II-701, IV-499, IV-549
 Yi, Subong III-144
 Yildiz, İpek I-547
 Yim, Keun Soo I-1000
 Yim, Soon-Bin II-1121
 Yoe, Hyun III-251
 Yoh, Jack Jai-ick V-484
 Yoo, Cheol-Jung III-188, III-222,
 IV-893, IV-955, V-644
 Yoo, Chuck II-1189
 Yoo, Chun-Sik V-958, V-979, V-990
 Yoo, Giljong II-430
 Yoo, Hwan-Hee V-1010
 Yoo, Hyeong Seon III-206, III-213
 Yoo, Jeong-Joon I-1000
 Yoo, Kee-Young I-1156, V-276, V-303
 Yoo, Ki-Sung II-1169
 Yoo, Kook-Yeol I-298
 Yoo, Sang Bong I-895
 Yoo, Seung Hwan II-270
 Yoo, Seung-Jae II-1025
 Yoo, Seung-Wha II-186
 Yoo, Sun K. I-335
 Yoon, Eun-Jun I-1156, V-276, V-303
 Yoon, Heejun II-11
 Yoon, Hwamook II-11
 Yoon, Kyunghyun I-402
 Yoon, Won Jin II-856
 Yoon, Won-Sik II-847
 Yoon, Yeo-Ran II-534
 Yoshizawa, Shuji I-596
 You, Ilsun IV-336, IV-416
 You, Young-Hwan V-752
 Youn, Hee Yong II-691, III-852,
 IV-1, IV-187, IV-456, IV-519,
 V-139, V-185
 Young, Chung Min II-534
 Yu, Jonas C.P. III-885
 Yu, Jun IV-649

Yu, Ki-Sung II-525, II-982
Yu, Lean I-518
Yu, Mei I-307, I-317
Yu, Sunjin II-87, II-96
Yu, Tae Kwon II-451
Yu, Young Jung I-904
Yu, Yung H. V-932
Yuen, Tsz Hon I-383
Yun, Jae-Kwan II-259
Yun, Kong-Hyun II-989

Zeng, Weiming II-964
Zhai, Jia IV-296
Zhang, David I-528
Zhang, Fan II-59
Zhang, Fanguo III-345
Zhang, Jianhong IV-262

Zhang, JianYu IV-262
Zhang, Jie V-149
Zhang, Minghu IV-529
Zhang, Xinhong II-59
Zhang, Zhongmei I-921
Zhao, Mingxi I-1118
Zheng, He II-954
Zheng, Lei IV-1090
Zheng, Nenggan II-1149
Zhiyong, Feng V-32
Zhong, Jingwei V-224
Zhou, Bo IV-352
Zhou, Yanmiao II-1149
Ziaee, M. III-574
Zongming, Wang I-885
Zou, Wei IV-262
Żyliński, Paweł I-141, I-161

ERRATUM

A Security Requirement Management Database Based on ISO/IEC 15408

Shoichi Morimoto, Daisuke Horie, and Jingde Cheng

Department of Information and Computer Sciences, Saitama University,
Saitama, 338-8570, Japan
{morimo, horie, cheng}@aise.ics.saitama-u.ac.jp

M. Gavrilova et al. (Eds.): ICCSA 2006, LNCS 3982, pp. 1–10, 2006.
© Springer-Verlag Berlin Heidelberg 2006

DOI 10.1007/11751595_1

An error has been found in the above article

- 1 In the original version of this paper the affiliation was not correct. The correct affiliation of Shoichi Morimoto, Daisuke Horie, and Jingde Cheng is : Department of Information and Computer Sciences, Saitama University, 338-8570, Japan.

The original online version for this chapter can be found at
http://dx.doi.org/10.1007/11751595_1
