Khabib Mustofa Erich J. Neuhold
A Min Tjoa Edgar Weippl
Ilsun You (Eds.)

LNCS 7804

Information and Communication Technology

International Conference, ICT-EurAsia 2013
Yogyakarta, Indonesia, March 2013
Proceedings

ifip

Lecture Notes in Computer Science 7804

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Khabib Mustofa Erich J. Neuhold
A Min Tjoa Edgar Weippl Ilsun You (Eds.)

Information and Communication Technology

International Conference, ICT-EurAsia 2013
Yogyakarta, Indonesia, March 25-29, 2013
Proceedings

Springer

Volume Editors

Khabib Mustofa
Universitas Gadjah Mada, Department of Computer Science and Electronics
Sekip Utara Bls, Yogyakarta 55281, Indonesia
E-mail: khabib@ugm.ac.id

Erich J. Neuhold
University of Vienna, Faculty of Computer Science
Währinger Straße 29, 1190 Wien, Austria
E-mail: erich.neuhold@univie.ac.at

A Min Tjoa
Vienna University of Technology
Institute of Software Technology and Interactive Systems
Favoritenstraße 9-11, 1040 Wien, Austria
E-mail: amin@ifs.tuwien.ac.at

Edgar Weippl
Vienna University of Technology and SBA Research
Institute of Software Technology and Interactive Systems
Favoritenstraße 9-11, 1040 Wien, Austria
E-mail: edgar.weippl@tuwien.ac.at

Ilsun You
Korean Bible University, School of Informatic Science
16 Danghyun 2-gil, Nowon-gu, Seoul 139-791, South Korea
E-mail: isyou@bible.ac.kr

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-36817-2 e-ISBN 978-3-642-36818-9
DOI 10.1007/978-3-642-36818-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013931976

CR Subject Classification (1998): K.4.1-4, K.6.5, H.3.4-5, H.4.1-3, H.5.1, H.5.3, E.3, K.3.1, C.2.0-5, D.2.0, I.2.7, D.2.4, C.5.3

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The ICT-EurAsia conference is thought as a platform for the exchange of ideas, experiences, and opinions among theoreticians and practitioners and for defining requirements of future systems in the area of ICT with a special focus on fostering long-term relationships among and with researchers and leading organizations in the Eurasian continent.

On the one hand, organizing cross-domain scientific events like this originated from an idea of IFIP President Leon Strous at the IFIP 2010 World Computer Congress in Brisbane and, on the other hand, by the many activities of the ASEA-UNINET (ASEAN-European Academic University Network) in the area. This network was founded in 1994 especially to enhance scientific and research cooperation between ASEAN countries and Europe. The great success of ASEA-UNINET led to the foundation of the EPU network (Eurasia Pacific University Network), which complements the geographic area of ASEA-UNINET covering the whole Eurasian continent. Both university networks have a strong focus on ICT.

The IFIP organizers of this event who are engaged within IFIP in the area of TC 5 (IT Applications) and TC 8 (Information Systems) also very much welcome the fertilization of this event by the collocation of AsiaARES as a special track on availability, reliability, and security. AsiaARES brings the idea of seven successful consecutive ARES conferences to Asia. ARES is devoted to building scientific bridges between the various aspects of system dependability and security as an integrated concept.

AsiaARES is a new conference track that specifically aims to provide better access to current IT-security research results for the Asian region**.** The ultimate goal is to establish a community and a meeting point for security researchers and professionals and to make travel shorter and the venues easily accessible for researchers from Asia. At AsiaARES we offer virtual presentations for which authors submit a video presentation in addition to their paper and the paper is discussed using online collaboration tools. This allows authors who have problems obtaining visas or very limited travel budgets to contribute.

We look forward to continuing the success of AsiaARES 2013 in further editions and improve AsiaARES based on the feedback of the authors and attendees – both virtual and real. Special thanks to Ilsun You for all his efforts toward this initiative.

We would like to express our thanks to all institutions actively supporting this event, namely:

- University Gadjah Mada, Yogyakarta, Indonesia
- International Federation for Information Processing (IFIP)
- ASEAN-European University Network
- Eurasia-Pacific University Network

- The Austrian Compentence Centres for Excellent Technology SBA (Secure Business Austria)
- The Austrian Agency for International Cooperation in Education and Research
- The Austrian Embassy in Jakarta and Its Honorary Consulate in Yogyakarta

The papers presented at this conference were selected after extensive reviews by the Program Committee (PC) and associated reviewers. We would like to thank all the PC members and the reviewers for their valuable advice, and the authors for their contributions.

Many persons contributed numerous hours to organize this conference. Their names appear on the following pages as committee members of this scientific conference.

Special thanks go to Amin Anjomshoaa, Ahmad Ashari, Muhammad Asfand-E-Yar, and last but not least to Yvonne Poul.

January 2013

Khabib Mustofa
Erich J. Neuhold
A Min Tjoa
Edgar Weippl
Ilsun You

Organization

Information & Communication Technology-EurAsia Conference 2013, ICT-EurAsia 2013

Honorary Committee

Abudul Jalil Nordin	Chairman of ASEA-UNINET
Atta Ur Rahman	President of the Academy of Science, Pakistan
Maria Raffai	Honorary Secretary IFIP
Roland Wagner	President of DEXA, Database and Expert Systems Association
Brigitte Winklehner	President EurAsia Pacific UNINET, EPU
Richardus Eko Indrajit	Chairman of APTIKOM, Association of Indonesian Universities on Informatics and Computer

General Chairs

Stephane Bressan	National University of Singapore, Singapore
Erich J. Neuhold	Chairman of IFIP Technical Committee on Information Technology Application

Program Committee Chairs

Ladjel Bellatreche	Laboratorie d'Informatique Scientifique et Industrielle, France
Alfredo Cuzzocrea	University of Calabria, Italy
Tran Khanh Dang	National University of Ho Chi Minh City, Vietnam
Mukesh Mohania	IBM Research India
Zhiyong Peng	Wuhan University, China
A Min Tjoa	Vienna University of Technology, Austria

Workshop Chairs

Andreas Holzinger	University of Graz, Austria
Ilsun You	Korean Bible University, South Korea

Panel Chairs

Inggriani Liem	ITB-Institute of Technology Bandung, Indonesia
Josef Küng	University of Linz, Austria
Lida Xu	Old Dominion University, USA

Organizing Committee

Saiful Akbar	Institute of Technology Bandung, Bandung
Ahmad Ashari	Universitas Gadjah Mada, Yogyakarta
Harno Dwi Pranowo	Universitas Gadjah Mada, Yogyakarta
M. Isa Irawan	Institute of Technology Sepuluh November, Surabaya
Khabib Mustofa	Universitas Gadjah Mada, Yogyakarta
Mardhani Riasetiawan	Universitas Gadjah Mada, Yogyakarta
Insap Santosa	Universitas Gadjah Mada, Yogyakarta
Widyawan	Universitas Gadjah Mada, Yogyakarta

Steering Committee

Masatoshi Arikawa	University of Tokyo, Japan
Wichian Chutimaskul	King Mongkut's University of Technology Thonburi, Thailand
Zainal A. Hasibuan	Universitas Indonesia, Indonesia
Hoang Huu Hanh	University of Hue, Vietnam
Jazi Eko Istiyanto	Universitas Gadjah Mada, Indonesia
Josef Küng	University of Linz, Austria
Ismail Khalil	Int. Org. for Information Integration and Web-based App. & Services
Inggriani Liem	Institute of Technology Bandung, Indonesia
Made Sudiana Mahendra	Udayana University, Indonesia
Pavol Navrat	Slovak University of Technology Bratislava, Slowakia
Günther Pernul	University of Regensburg, Germany
Maria Raffai	University of Györ, Hungary
Sri Wahyuni	Universitas Gadjah Mada, Indonesia

Organizational Coordination Chair

Muhammad Asfand-e-yar	Vienna University of Technology, Austria

Senior Program Committee

Silvia Abrahao	University Politecnica de Valencia, Spain
Hamideh Afsarmanesh	University of Amsterdam, Netherlands
Masatoshi Arikawa	University of Tokyo, Japan
Hyerim Bae	Pusan National University, Korea
Sourav S. Bhowmick	Nanyang Technological University, Singapore
Nguyen Thah Binh	IIASA, Austria
Robert P. Biuk-Aghai	University of Macau, China

Manfred Broy	Technical University of Munich, Germany
Indra Budi	University of Indonesia, Indonesia
Gerhard Budin	University of Vienna, Austria
Somchai Chatvichienchai	University of Nagasaki, Japan
Key Sun Choi	KAIST, Korea
Stavros Christodoulakis	Technical University of Crete, Greece
Wichian Chutimaskul	KMUTT, Thailand
Hoang Xuan Dau	PTIT, Hanoi, Vietnam
Duong Anh Duc	University of Information Technology, Vietnam
Tetsuya Furukawa	University of Kyushu, Japan
Andrzej Gospodarowicz	Wroclaw University of Economics, Poland
Zainal Hasibuan	University of Indonesia, Indonesia
Christian Huemer	Vienna University of Technology, Austria
Mizuho Iwaihara	Faculty of Science and Engineering Waseda University, Japan
Matthias Jarke	RWTH Aachen Lehrstuhl Informatik, Germany
Gerti Kappel	Vienna University of Technology, Austria
Dimitris Karagiannis	University of Vienna, Austria
Shuaib Karim	Quaid-i-Azam University, Pakistan
Sokratis Katsikas	University of Piraeus, Greece
Dieter Kranzlmüller	Ludwig-Maximilians-Universität München, Germany
Narayanan Kulathuramaiyer	Universiti Malaysia Sarawak, Malaysia
Josef Küng	Johannes Kepler Universität Linz, Austria
Khalid Latif	National University of Sciences and Technolgy, Pakistan
Lenka Lhotska	Czech Technical University, Czech Republic
Inggriani Liem	ITB-Institute of Technology Bandung, Indonesia
Peri Loucopoulos	Loughborough University, UK
Vladimir Marik	Czech Technical University, Czech Republic
Luis M. Camarinha Matos	Universidade Nova de Lisboa, Portugal
Günter Müller	University of Freiburg, Germany
Thoai Nam	HCMC University of Technology, Vietnam
Günther Pernul	University of Regensburg, Germany
Geert Poels	Ghent University, Belgium
Gerald Quirchmayr	University of Vienna, Austria
Dana Indra Sensuse	University of Indonesia, Indonesia
Josaphat Tetuko Sri Sumantyo	Chiba University, Japan
Wikan Danar Sunindyo	Institute of Technology Bandung, Indonesia
Katsumi Tanaka	Kyoto University, Japan
Juan Trujillo	University of Alicante, Spain
Nguyen Tuan	Vietnam National University, Vietnam
Werner Winiwarter	University of Vienna, Austria

The 2013 Asian Conference on Availability, Reliability and Security, AsiaARES 2013

Program Committee Chair

Ilsun You	Korean Bible University, South Korea

Program Committee

Dong Seong Kim	University of Canterbury, New Zealand
Kyung-Hyune Rhee	Pukyong National University, Republic of Korea
Kangbin Yim	Soonchunhyang University, Republic of Korea
Qin Xin	University of the Faroe Islands, Denmark
Pandu Rangan Chandrasekaran	Indian Institute of Technology Madras, India
Shinsaku Kiyomoto	KDDI R&D Laboratories Inc., Japan
Atsuko Miyaji	JAIST, Japan
Willy Susilo	University of Wollongong, Australia
Xiaofeng Chen	Xidian University, China
Shuichiroh Yamamoto	Nagoya University, Japan
Fangguo Zhang	Sun Yan-Sen Univerisity, China
Xinyi Huang	Fujian Normal University, China
Qianhong Wu	Universitat Rovira i Virgili, Spain
Zhang Jie	Nanyang Technological University, Singapore
Rana Barua	Indian Statistical Institute, India
Baokang Zhao	National University of Defense Technology, China
Joonsang Baek	Khalifa University of Science, Technology & Research, Kustar, UAE
Fang-Yie Leu	Tunghai University, Taiwan
Francesco Palmieri	Seconda Università di Napoli, Italy
Aniello Castiglione	Università degli Studi di Salerno, Italy
Ugo Fiore	Seconda Università di Napoli, Italy
Kouichi Sakurai	Kyushu University, Japan
Yizhi Ren	Hangzhou Dianzi University, China
Sushmita Ruj	Indian Institute of Technology, IIT-Indore, India
Kirill Morozov	Kyushu University, Japan
Chunhua Su	Institute for Infocomm Research, Singapore
Ren Junn Hwang	Tamkang University, Taiwan
Shiuh-Jeng Wang	Central Police University, Taiwan

Table of Contents

Information and Communication Technology-Eurasia Conference (ICT-EurAsia)

E-Society

Translating the Idea of the eGovernment One-Stop-Shop in Indonesia 1
Fathul Wahid

A Practical Solution against Corrupted Parties and Coercers in Electronic Voting Protocol over the Network 11
Thi Ai Thao Nguyen and Tran Khanh Dang

Early-Detection System for Cross-Language (Translated) Plagiarism 21
Khabib Mustofa and Yosua Albert Sir

TransWiki: Supporting Translation Teaching 31
Robert P. Biuk-Aghai and Hari Venkatesan

Physicians' Adoption of Electronic Medical Records: Model Development Using Ability – Motivation - Opportunity Framework 41
Rajesri Govindaraju, Aulia F. Hadining, and Dissa R. Chandra

Software Engineering

Software Development Methods in the Internet of Things 50
Selo Sulistyo

SAT–Based Bounded Strong Satisfiability Checking of Reactive System Specifications 60
Masaya Shimakawa, Shigeki Hagihara, and Naoki Yonezaki

OSMF: A Framework for OSS Process Measurement 71
Wikan Danar Sunindyo and Fajar Juang Ekaputra

Analyzing Stability of Algorithmic Systems Using Algebraic Constructs 81
Susmit Bagchi

Algorithms of the Combination of Compiler Optimization Options for Automatic Performance Tuning 91
Suprapto and Retantyo Wardoyo

Security and Privacy

On Efficient Processing of Complicated Cloaked Region for Location Privacy Aware Nearest-Neighbor Queries 101
Chan Nam Ngo and Tran Khanh Dang

Semantic-Aware Obfuscation for Location Privacy at Database Level . . . 111
Thu Thi Bao Le and Tran Khanh Dang

Practical Construction of Face-Based Authentication Systems with Template Protection Using Secure Sketch 121
Tran Tri Dang, Quynh Chi Truong, and Tran Khanh Dang

CAPTCHA Suitable for Smartphones 131
Yusuke Tsuruta, Mayumi Takaya, and Akihiro Yamamura

Code Based KPD Scheme with Full Connectivity: Deterministic Merging 141
Pinaki Sarkar and Aritra Dhar

Cloud and Internet Computing

Indonesian Digital Natives: ICT Usage Pattern Study across Different Age Groups 152
Neila Ramdhani and Wisnu Wiradhany

End-to-End Delay Performance for VoIP on LTE System in Access Network 162
Liang Shen Ng, Noraniah Abdul Aziz, and Tutut Herawan

Mobile Collaboration Technology in Engineering Asset Maintenance – What Technology, Organisation and People Approaches Are Required? 173
Faisal Syafar and Jing Gao

A Genetic Algorithm for Power-Aware Virtual Machine Allocation in Private Cloud 183
Quang-Hung Nguyen, Pham Dac Nien, Nguyen Hoai Nam, Nguyen Huynh Tuong, and Nam Thoai

Cloud-Based E-Learning: A Proposed Model and Benefits by Using E-Learning Based on Cloud Computing for Educational Institution 192
Nungki Selviandro and Zainal Arifin Hasibuan

Knowledge Management

Information Systems Strategic Planning for a Naval Hospital 202
Hery Harjono Muljo and Bens Pardamean

Estimation of Precipitable Water Vapor Using an Adaptive Neuro-fuzzy Inference System Technique 214
Wayan Suparta and Kemal Maulana Alhasa

A Data-Driven Approach toward Building Dynamic Ontology 223
Dhomas Hatta Fudholi, Wenny Rahayu, Eric Pardede, and Hendrik

Using Semantic Web to Enhance User Understandability for Online Shopping License Agreement 233
Muhammad Asfand-e-yar and A Min Tjoa

Asian Conference on Availability, Reliability and Security (AsiaARES)

Dependable Systems and Applications

Secure and Verifiable Outsourcing of Sequence Comparisons 243
Yansheng Feng, Hua Ma, Xiaofeng Chen, and Hui Zhu

Syntactic Analysis for Monitoring Personal Information Leakage on Social Network Services: A Case Study on Twitter 253
Dongjin Choi, Ilsun You, and Pankoo Kim

On the Efficiency Modelling of Cryptographic Protocols by Means of the Quality of Protection Modelling Language (QoP-ML) 261
Bogdan Ksiezopolski, Damian Rusinek, and Adam Wierzbicki

DiffSig: Resource Differentiation Based Malware Behavioral Concise Signature Generation 271
Huabiao Lu, Baokang Zhao, Xiaofeng Wang, and Jinshu Su

On Identifying Proper Security Mechanisms 285
Jakub Breier and Ladislav Hudec

A Recovery Approach for SQLite History Recorders from YAFFS2 295
Beibei Wu, Ming Xu, Haiping Zhang, Jian Xu, Yizhi Ren, and Ning Zheng

UVHM: Model Checking Based Formal Analysis Scheme for Hypervisors 300
Yuchao She, Hui Li, and Hui Zhu

SA4WSs: A Security Architecture for Web Services 306
Lingxia Liu, Dongxia Wang, Jinjing Zhao, and Minhuan Huang

Verifying Data Authenticity and Integrity in Server-Aided Confidential Forensic Investigation 312
Shuhui Hou, Ryoichi Sasaki, Tetsutaro Uehara, and Siuming Yiu

A Test Case Generation Technique for VMM Fuzzing 318
Xiaoxia Sun, Hua Chen, Jinjing Zhao, and Minhuan Huang

A Variant of Non-Adaptive Group Testing and Its Application in Pay-Television via Internet 324
Thach V. Bui, Oanh K. Nguyen, Van H. Dang, Nhung T.H. Nguyen, and Thuc D. Nguyen

A Proposal on Security Case Based on Common Criteria 331
Shuichiro Yamamoto, Tomoko Kaneko, and Hidehiko Tanaka

An Adaptive Low-Overhead Mechanism for Dependable General-Purpose Many-Core Processors 337
Wentao Jia, Rui Li, and Chunyan Zhang

Identity Management Lifecycle - Exemplifying the Need for Holistic Identity Assurance Frameworks 343
Jostein Jensen

Cryptography

Anonymous Lattice-Based Broadcast Encryption 353
Adela Georgescu

Supporting Secure Provenance Update by Keeping "Provenance" of the Provenance 363
Amril Syalim, Takashi Nishide, and Kouichi Sakurai

New Ciphertext-Policy Attribute-Based Access Control with Efficient Revocation 373
Xingxing Xie, Hua Ma, Jin Li, and Xiaofeng Chen

Provably Secure and Subliminal-Free Variant of Schnorr Signature 383
Yinghui Zhang, Hui Li, Xiaoqing Li, and Hui Zhu

A Block Cipher Mode of Operation with Two Keys 392
Yi-Li Huang, Fang-Yie Leu, Jung-Chun Liu, and Jing-Hao Yang

On the Security of an Authenticated Group Key Transfer Protocol Based on Secret Sharing 399
Ruxandra F. Olimid

Modified Efficient and Secure Dynamic ID-Based User Authentication Scheme 409
Toan-Thinh Truong, Minh-Triet Tran, and Anh-Duc Duong

Privacy and Trust Management

A Simplified Privacy Preserving Message Delivery Protocol in VDTNs 416
Youngho Park, Chul Sur, and Kyung-Hyune Rhee

Confidentiality-Preserving Query Execution of Fragmented Outsourced Data 426
Anis Bkakria, Frédéric Cuppens, Nora Cuppens-Boulahia, and José M. Fernandez

Enhancing Privacy Protection in Distributed Environments through Identification and Authentication-Based Secure Data-Level Access Control 441
Nisreen Alam Aldeen and Gerald Quirchmayr

Toward Secure Clustered Multi-Party Computation: A Privacy-Preserving Clustering Protocol 447
Sedigheh Abbasi, Stelvio Cimato, and Ernesto Damiani

A Real-Time Privacy Amplification Scheme in Quantum Key Distribution 453
Bo Liu, Bo Liu, Baokang Zhao, Dingjie Zou, Chunqing Wu, Wanrong Yu, and Ilsun You

Network Analysis and Security

CSP-Based General Detection Model of Network Covert Storage Channels 459
Hui Zhu, Tingting Liu, Guanghui Wei, Beishui Liu, and Hui Li

Trustworthy Opportunistic Access to the Internet of Services 469
Alessandro Armando, Aniello Castiglione, Gabriele Costa, Ugo Fiore, Alessio Merlo, Luca Verderame, and Ilsun You

Architecture of Network Environment for High-Risk Security Experimentation 479
Xiaohui Kuang, Xiang Li, and Jinjing Zhao

Emulation on the Internet Prefix Hijacking Attack Impaction 485
Jinjing Zhao and Yan Wen

Improved Clustering for Intrusion Detection by Principal Component Analysis with Effective Noise Reduction 490
Lu Zhao, Ho-Seok Kang, and Sung-Ryul Kim

Unconditionally Secure Fully Connected Key Establishment Using Deployment Knowledge 496
Sarbari Mitra, Sourav Mukhopadhyay, and Ratna Dutta

An Improved Greedy Forwarding Routing Protocol for Cooperative VANETs 502
Huaqing Wen and Kyung-Hyune Rhee

A Review of Security Attacks on the GSM Standard 507
Giuseppe Cattaneo, Giancarlo De Maio, Pompeo Faruolo, and Umberto Ferraro Petrillo

Multimedia Security

An Extended Multi-secret Images Sharing Scheme Based on Boolean Operation 513
Huan Wang, Mingxing He, and Xiao Li

Image Watermarking Using Psychovisual Threshold over the Edge 519
Nur Azman Abu, Ferda Ernawan, Nanna Suryana, and Shahrin Sahib

A Data Structure for Efficient Biometric Identification 528
Kensuke Baba and Serina Egawa

The PCA-Based Long Distance Face Recognition Using Multiple Distance Training Images for Intelligent Surveillance System 534
Hae-Min Moon and Sung Bum Pan

Shifting Primes on OpenRISC Processors with Hardware Multiplier 540
Leandro Marin, Antonio J. Jara, and Antonio Skarmeta

Author Index 551

Translating the Idea of the eGovernment One-Stop-Shop in Indonesia

Fathul Wahid

Department of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia
Department of Information Systems, University of Agder, Kristiansand, Norway
fathul.wahid@uii.ac.id

Abstract. This study aims to understand how the idea of an eGovernment one-stop-shop (OSS) has been translated into a new setting. Since the beginning of 2000, this idea has been implemented in a variety of ways by Indonesian local governments. Using an interpretive case study in the city of Yogyakarta, the study revealed that the specificity of each setting influences the translation process of the idea of OSS during its institutionalization. It also identified a set of editing rules used during the translation process. These include the editing rules concerning context (e.g., internal readiness); logic (e.g., corruption eradication); and formulation (e.g., standardized processes). The study also found that the idea translation was not a single round process.

Keywords: one-stop-shop, eGovernment, idea translation, editing rules, institutionalization, institutional theory, developing countries, Indonesia.

1 Introduction

In its 2010 survey, the Political and Economic Risk Consultancy named Indonesia as one of the Asia's most inefficient bureaucracies that downgraded the quality of public services and discouraged investment [1]. This was a general assessment and did not provide detailed insights. If we scrutinize the state of bureaucracy at the local government level (i.e., city/district), a different picture emerges. For example, the World Bank [2] placed the city of Yogyakarta fifth in a list of the most efficient bureaucracies in terms of dealing with construction licenses among 183 economies in a global survey. In short, there is a huge discrepancy between different local governments in the quality of public services they provide [2, 3].

In order to improve the quality of public service and eradicate corruption, the government of Indonesia has taken various initiatives. At the local government level, one of these initiatives was translated into the establishment of an eGovernment one-stop-shop (OSS). In this paper, OSS refers to the licensing department that provides a variety of services (such as construction and nuisance licenses) to the public. With the help of information technology (IT), the OSS was designed to cut red tape, 'burdensome administrative rules and procedures' [4:385].

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 1–10, 2013.

However, not all OSSs in Indonesia have successfully achieved their potential for providing a better public service and curbing corrupt practices [3, 5]. Despite this, some local governments have enjoyed benefits from the successful implementation of the initiative. The city of Yogyakarta is one of them. This study aims to explain how the idea of OSS is translated in the institutionalization process at the context of local government. It is also motivated by a lack of studies that pay attention to such the process in the eGovernment initiative implementation [6, 7]. Hence, the main research question addressed by this study is: *how is the idea of an eGovernment one-stop shop translated during its institutionalization process*? The concepts of idea translation introduced by the Scandinavian institutionalism are used to explain the process of OSS institutionalization [8].

2 Theoretical Framework

2.1 eGovernment One-Stop-Shop

The establishment of OSS that can be considered as an implementation of 'joined-up government' [9, 10], may provide four benefits: eliminating the contradictions and tensions between different policies, making better use of scarce resources, improving the flow of good ideas and synergy between different stakeholders, and creating seamless rather than fragmented services. In the context of developing countries, the establishment of OSS is very important in cutting red tape and eradicating corruption [11, 12]. The use of IT can help to reduce hierarchical structures and streamline the process of filtering out unnecessary impediments to efficient operation [4].

However, there are two inherent problems in this regard: a problem of coordination and a problem of integration and organization [13]. The former involves encouraging the agencies involved to work on broadly the same agenda, while the latter concerns the problem of how to align structures, incentives, and cultures to fit inter-organizational tasks. Both of these are institutional problems. Previous studies argue that eGovernment will not achieve its potential without institutional change [14, 15].

Thus, in the context of Indonesia, the OSS has not yet really been able to provide effective online services. The citizens may get information, downloadable forms, and trace the status of an application from a website, but they cannot send the application online. In order to do so they have to visit the OSS physically to hand in applications and to make payments. In this context, we may consider the OSS as an eGovernment 'official' intermediary[1] that helps citizens to get the public services they need. The role of intermediaries in providing eGovernment services in the context of developing countries is very influential [16].

[1] This term is used to differentiate between 'official' and 'unofficial' intermediaries. OSS is a manifestation of the former, while the latter are often not immune from corrupt practices (i.e., petty bribery). A survey conducted in Indonesia found that 48% of license applicants used the 'unofficial' intermediaries (i.e., local government staffs), and the use of such intermediaries increased the licensing costs by 58%, despite the fact that this speeded up the process [3].

2.2 Idea Translation

The concepts of idea translation originated from the Scandinavian institutional research can be said to "primarily come to highlight the dynamic aspect of circulating ideas; how and why ideas become wide-spread, how they are translated as they flow and with what organization consequences" [17:219]. Answers to these questions are needed to explain the process of translation of an idea in a certain setting. An organization does not operate in a vacuum, and an idea that is picked up can be adopted and incorporated into organizational practice. When an idea is adopted, it does not always work as planned and in many cases it can then be decoupled from ongoing activities of organization [18]. In order to work in its new setting, the idea requires a process of translation [19].

The translation process is described by Czarniawska and Joerges [20] in four stages: idea, object, action and institution. In a particular context, organizational actors select an idea among a collection of circulating ideas. The circulating ideas are disembedded from their original setting, before being reembedded into a new setting [20]. Once an idea is chosen, it will be subsequently transformed into an object. The objectification process makes the idea tangible. The easiest way to objectify ideas is to turn them into linguistic artefacts, such as labels and metaphors [20]. An idea can then be translated into an object (e.g. a prototype, text, model, perceptions, a concept) and can then be realised. Thus, an object becomes translated into an action. Finally, it may emerge as an institution if the action is regularly repeated over time and therefore becomes taken for granted.

However, the process of translation of an idea is in fact constrained by the editing rules which are often implicitly inherent within an organization [17]. In general the editing rules concern three factors: context, logic, and formulation [19]. Different settings may follow a variable set of idea editing rules. Ideas may be contextualized to consider aspects of time, space, and scale. New types of logic or explanations can be accepted; and/or be formulated as, e.g., a prototype to attract attention. The idea translation process is depicted schematically in Fig. 1.

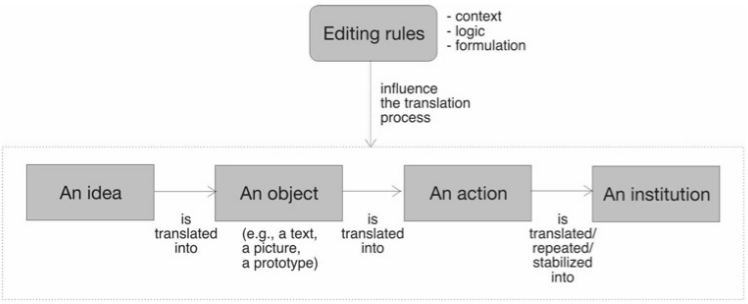


Fig. 1. The process of idea translation (adopted from [20])

3 Research Method

This case study is interpretative in nature. In choosing this approach, "our knowledge of reality is gained only through social constructions such as language, consciousness, shared meanings, documents, tools, and other artefacts" [21:69]. A case study is appropriate since the problem under investigation is practice-based, the experiences of the actors are important and the context of action is critical [22].

Data were collected mainly through interviews. Interviews were carried out with a variety of key players with OSS and/or eGovernment implementation at various levels. The interviewees were the mayor of the city of Yogyakarta, four heads/vice-heads of offices, three heads of divisions and one administrator. The snowball method was used to select the interviewees.

Eight interviews were conducted, most of which were recorded. Each interview lasted 30 to 60 minutes. The interviews were conducted between July and August 2011. To ensure the validity of the data [23], additional data were also collected from written documents/reports and field observations. The concepts of idea translation were used as templates when coding the data, and temporal bracketing as sensemaking strategy [24] was used in the data analysis.

4 Findings

4.1 OSS as a Unit with a Limited Authority: 2000-2001

The OSS in Yogyakarta can be traced back to 2000, when the local government decided to establish a one-roof service unit (*Unit Pelayanan Terpadu Satu Atap* [UPTSA]) to response the national regulation. This was the origin of the idea of the OSS. However, the regulation did not provide a comprehensive guide on how the OSS should be established.

At that time, UPTSA acted as a front office counter for 12 services that *received* the applications, whereby the mechanism to process the applications was similar to those that existed before its establishment. As *a unit*, UPTSA had no authority to approve the applications, but passed them on to the technical department that had the authority to issue the licenses. Thus, in the process of getting a license, the burdensome bureaucracy continued, although with some subtle improvements. A number of specific problems had to be coped with, such as the state of **internal readiness**. One informant asserted that:

"At that time, there was a lack of political will both from the mayor and the heads of departments. … [There was] no independent budget allocation and no independent institution. We were not ready for that time."

The status of UPTSA also made smooth coordination with other departments difficult, as there was lack of integration between them.

A momentum emerged when a new mayor was elected in 2001. Before taking on the position, the mayor had been a successful businessman. He was interested on

UPTSA, since he had experienced difficulties in the past when dealing with burdensome bureaucracy. The new mayor showed his political will to improve UPTSA. A comprehensive evaluation was carried out, involving all departments as part of the preparation process. The former head of UPTSA stated that:

"So, in 2001, we identified the authorities of all departments and simplified them. … We identified which the licenses that could be integrated into the licensing department. Some of the licenses were very specific such as those for medical doctors and nurses. These kinds of licenses were still the authority of the respective technical departments."

4.2 OSS as a Unit with a Higher Authority: 2002-2005

Based on a set of recommendations from a comprehensive evaluation, some improvements were made. Among the problems identified at that time was the lack of a smooth flow of service provision, the need to assign employees to tasks effectively, to appoint a coordinator for UPTSA, and to improve the supporting facilities. In 2002, the mayor re-launched UPTSA. This **re-launching initiative** created a new momentum. One informant asserted that:

"In January 2002, supporting facilities of the UPTSA office were improved. … UPTSA was then re-launched on March 4, 2002 by the mayor."

Since then, UPTSA has had its own budget. At that time the status of the officers who worked for it was still attached to the departments they had originated from. However, UPTSA included the possibility of *coordinating* the processing of the applications. Thus, the problems of coordination and integration were partly solved.

In 2003, a Government Regulation (*Peraturan Pemerintah*) No. 8/2003 concerning Guidelines for Local Government Organization was enacted. This **national regulation** made a **comprehensive reorganization** of all local government agencies possible. The existence of **power interplay** between actors was apparent at this stage. After an evaluation process, it was agreed that UPTSA should become a definitive government agency in the near future. The former head of UPTSA stated:

"When there was a new regulation from the national government, we did not take it for granted. … We had to think holistically about organizational structure, personnel, budgeting, and authority. Taking away the authority [from a department] was not easy."

It was agreed to promote the status of UPTSA from a unit into a department (*dinas*), which legally would have a higher authority. The decision was was also a result of the idea translation process. At that time, the mayor showed his **political leadership**, by asking all the heads of department who did not agree with the decision to express their opinion. As one informant stated:

"The mayor invited all the heads of department involved. … The mayor asked whom did not agree with the idea [of establishing an OSS as a department] to sign a statement on a paper bearing IDR 6,000 duty stamp. No one did it."

There was also a shift in the role of IT in supporting licensing services. As one informant asserted:

"In 2000, the information technology section was just a supporting unit. … But after making an organizational evaluation, the section became a core section. It happened in 2003."

Starting from 2003, **several information systems (IS) were developed** to support application processing. The use of IS helped to **standardize the application process**. Since that time, the application forms have been available online and a call center has been opened to provide information and to collect input and complaints from the citizens. A **continuous evaluation** procedure asked all the applicants to fill in a questionnaire to assess the quality of service from various points of view.

In 2004 the OSS initiative then received more impetus when the national government of Indonesia asked all government agencies to improve the quality of public service delivery, through **Presidential Instruction** No. 5/2004, as part of an effort to eradicate corruption. Then, the **corruption eradication** became **institutional logic** behind the establishment of OSS. The former head of OSS stated that:

"What we restructured at that time was not the licensing department, but all the departments. We identified what processes should be carried out by which department. No one complained, since we did not know to what department we would be assigned to."

4.3 OSS as a Department: 2006-Present

After going through an intricate process, at the end of 2005, through a local regulation (*Peraturan Daerah*) the mayor promoted the status of UPTSA as a unit to become a department. By using this new status, the Licensing Department (*Dinas Perizinan*) had the authority to process and *to approve/disapprove* applications. The department at that time had the responsibility for 35 types of licenses, while 24 other licenses were still being processed by the Health Department, due to their technical nature.

From the beginning, one of the main challenges was **a new culture building**, since the officers came from various departments. The (former) head of OSS cultivated the values of togetherness among the officers. It took around six months to build this new culture. In addition to the corruption eradication logic, there were also other **institutional logics** behind the OSS establishment and its development, from **public services** to **internal process improvement**. As one informant stated:

"The licensing department was established to improve the public services. We were on the public's side; the procedures were simplified. But, after the department was running, then I realized that the local government itself enjoyed the most advantages; including time efficiency, more controllable processing times, cost reductions and energy efficiency. No need to spend energy on coordination between the departments, since we were integrated."

The problems of integration and coordination were thus largely solved at this stage. To provide a better service, in 2007 a new organizational structure was adopted. Some procedures were **simplified**. In 2008 the number of licenses was reduced from 35 to

29. Some technical licenses, however, such as license for a medical doctor, were still under the authority of the related technical departments.

Since 2007, the licensing department obtained recognition from various national and international institutions. Such recognition included an Investment Award in 2007 and 2008 from the Investment Coordinating Board (*Badan Koordinasi Penanaman Modal* [BKPM]) and the Service Excellence Award in 2008 from the Ministry for the Empowerment of State Apparatus. In 2010, Yogyakarta was named as the fifth most efficient bureaucracy in dealing with construction licenses among 183 surveyed economies [2]. In 2011 it obtained an ISO 9001:2008 certificate for quality assurance.

5 Discussion

This discussion focused on from answering the research questions set at the outset: how is the idea of eGovernment one-stop shop translated during its institutionalization process? In doing so, the four-stage idea translation process [20] was central.

This study found that the process of idea translation was not one-way. This study revealed that the idea of OSS has been translated three times (in 2000, 2002, and 2006). Here, the object (i.e., the concept of OSS) has been evaluated through the process of shaping and are being shaped by everyday practices (i.e., action and/or institution). This finding provided new insights and a theoretical contribution to the concept of idea translation that seemed to be simplified as a one-way translation process [cf. 20]. Fig. 2 depicts the contextualized process of idea translation.

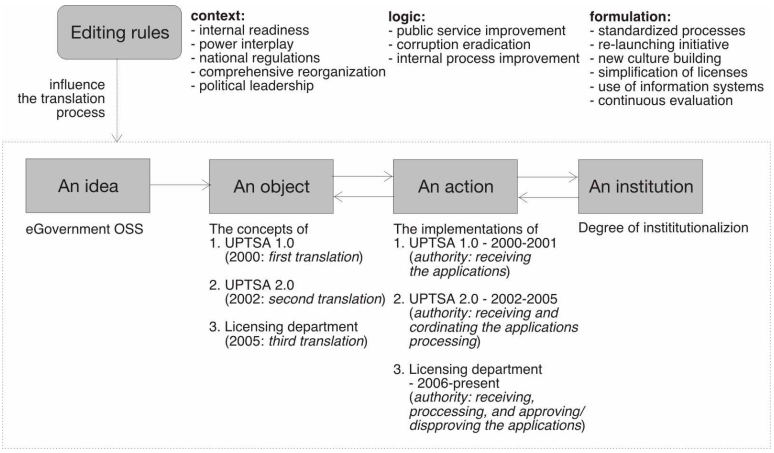
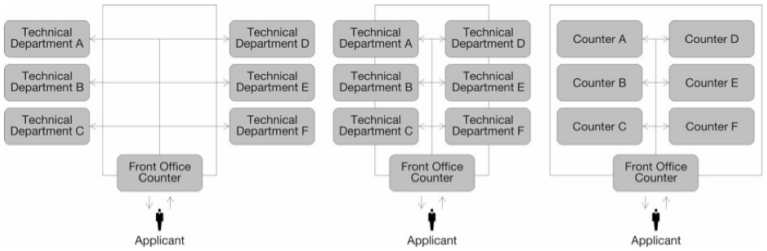


Fig. 2. The translation process of the OSS idea

Based on the presentation of the findings, several editing rules were identified during the process of idea translation (marked by a bold typeface in the Findings section) (see Fig 2.). Although the *idea* of OSS came from the national government, it did not provide detailed guidelines as to how an OSS should be

established. It is necessary to ask why no such guidelines were provided by the national government. Detailed guidelines can have both a negative and a positive impact. In considering the specificity of each local government, imposing detailed guidelines might be misleading and make localization impossible. However, for local governments with limited exposure to external knowledge and/or with many competing institutional logics within their organisations, such guidelines would be very useful. Otherwise, the idea could be self-defeating [25].

In 2000, the idea of OSS was translated into an object (i.e., the concept of UPTSA 1.0[2]) and then into an action (its implementation). At that time, the concept was not well developed, since there were problems of integration and coordination. UPTSA 1.0 as a unit had only limited authority. At this stage, the authority of OSS was only to receive the application (see Fig. 3(a)). A lack of political will was identified as one of main challenges at that time. Due to these problems of integration and coordination [13], the first idea translation failed to become institutionalized.



(a) (b) (c)

Fig. 3. Levels of OSS authority in Yogyakarta: (a) as a unit (2000-2001) – receiving; (b) as a unit (2002-2005) – receiving and coordinating; (c) as a department (2006-present) – receiving, processing, and approving. Note: The visual presentation was inspired by Steer [3].

The second attempt to translate the idea of OSS was carried out in 2002, when UPTSA 1.0 was re-launched as UPTSA 2.0. An evaluation report made at the beginning 2002 was the new translation object. The first translation was then corrected by improving the integration and coordination process between the involved departments. Since then, OSS has had a *coordinating* authority to process the application, in addition to just *receiving* it (see Fig. 3(b)). The problems of integration and coordination were *partly* solved. UPTSA 2.0 began to achieve a momentum towards becoming institutionalized by applying more standardized practices, supported by an improved IT infrastructure. However, there was still a need to streamline the application processing hindered by the limited authority of OSS.

At the third attempt, after carrying out comprehensive evaluation and preparation by the end of 2005, the idea of OSS was translated into a concept of a more integrated and a better-coordinated service provision. In 2006, UPTSA 2.0 became a licensing department with a higher authority (see Fig. 3(c)). This new status as a department

[2] The term 'UPTSA 1.0' is used to differentiate it from 'UPTSA 2.0' after a re-launching in 2002.

largely solved both the problems of integration and coordination. The new practice became routine through typification process where certain forms of actions came to be associated with certain classes of actors [26]. This process helped the licensing department become institutionalized.

6 Conclusion

This paper has presented the process of how an idea of OSS was translated in a specific setting. The concept of idea translation from Scandinavian institutionalism was used as a focus. This study has made two main contributions. Firstly, it has offered an explanation of how the same idea of OSS can be translated differently when it is implemented in a new setting. A new application of the concept of idea translation in the context of eGovernment studies can bring about a better understanding of the process of localization or local improvisation of an eGovernment initiative. A set of editing rules concerning context, logic, and formulation was also identified to explain this process. Secondly, theoretically, the study has offered empirical evidence to incorporate the multiple-round idea translation process.

This study was not without its limitations. It focused on a single case. Including various cases of the translation process of the idea of OSS, may reveal a more comprehensive picture of the possible process and its editing rules. However, as an interpretive study, the findings are generalized to theoretical concepts rather than the population [27]. As such we can make inferences about the concept of idea translation. It would also be interesting for future research to identify the circumstances in which the one-way or multiple-round translation process could be made more favourable.

References

1. Anonymous: India, Indonesia 'worst for red tape'. The Sydney Morning Herald (2010)
2. The World Bank and The International Finance Corporation: Doing Business in Indonesia 2010. The World Bank and The International Finance Corporation, Washington (2010)
3. Steer, L.: Business Licensing and One Stop Shops in Indonesia. The Asia Foundation (2006)
4. Welch, E.W., Pandey, S.K.: E-Government and bureaucracy: Toward a better understanding of intranet implementation and its effect on red tape. Journal of Public Administration Research and Theory 17, 379–404 (2006)
5. The Asia Foundation: Making Sense of Business Licensing in Indonesia: A Review of Business Licensing Policy and Survey of One Stop Shop Service Centers. The Asia Foundation, San Francisco (2007)
6. Wahid, F.: The Current State of Research on eGovernment in Developing Countries: A Literature Review. In: Scholl, H.J., Janssen, M., Wimmer, M.A., Moe, C.E., Flak, L.S. (eds.) EGOV 2012. LNCS, vol. 7443, pp. 1–12. Springer, Heidelberg (2012)
7. Wahid, F.: Themes of research on eGovernment in developing countries: Current map and future roadmap. In: Proceedings of the 46th Hawaii International Conference on System Sciences (HICSS 2013) (2013)

8. Czarniawska, B., Sevón, G.: Introduction. In: Czarniawska, B., Sevón, G. (eds.) Translating Organizational Change, pp. 1–12. Walter de Gruyter, Berlin (1996)
9. Pollitt, C.: Joined-up government: A survey. Political Studies Review 1, 34–49 (2003)
10. Persson, A., Goldkuhl, G.: Joined–Up E–Government – Needs and Options in Local Governments. In: Wimmer, M.A., Scholl, H.J., Janssen, M., Traunmüller, R. (eds.) EGOV 2009. LNCS, vol. 5693, pp. 76–87. Springer, Heidelberg (2009)
11. Guriev, S.: Red tape and corruption. Journal of Development Economics 73, 489–504 (2004)
12. Bardhan, P.: Corruption and development: a review of issues. Journal of Economic Literature 35, 1320–1347 (1997)
13. Mulgan, G.: Joined up government: Past, present and future. In: Bogdanor, V. (ed.) Joined-up Government. Oxford University Press, Oxford (2005)
14. Avgerou, C.: IT and organizational change: An institutionalist perspective. Information Technology and People 13, 234–262 (2000)
15. Furuholt, B., Wahid, F.: E-government challenges and the role of political leadership in Indonesia: The case of Sragen. In: Proceedings of the 41st Hawaii International Conference on System Sciences (HICSS 2008) (2008)
16. Sein, M.K., Furuholt, B.: Intermediaries: Bridges across the digital divide. Information Technologies for Development 18, 332–344 (2012)
17. Sahlin, K., Wedlin, L.: Circulating ideas: Imitation, translating and editing. In: Greenwood, R., Oliver, C., Suddaby, R., Sahlin, K. (eds.) The SAGE Handbook of Organizational Institutionalism, pp. 218–242. Sage, Los Angeles (2008)
18. Meyer, J.W., Rowan, B.: Institutional organizations: Formal structure as myth and ceremony. American Journal of Sociology 83, 340–463 (1977)
19. Sahlin-Andersson, K.: Imitating by editing success: The construction of organizational fields. In: Czarnlawska, B., Sevon, G. (eds.) Translating Organizational Change, pp. 69–92. Walter de Gruyter, Berlin (1996)
20. Czarniawska, B., Joerges, B.: Travel of ideas. In: Czarniawska, B., Sevón, G. (eds.) Translating Organizational Change, pp. 13–48. Walter de Gruyter, Berlin (1996)
21. Klein, H.K., Myers, M.D.: A set of principles for conducting and evaluating interpretive field studies in information systems. MIS Quarterly 23, 67–93 (1999)
22. Benbasat, I., Goldstein, D.K., Mead, M.: The case research strategy in studies of information systems. MIS Quarterly 11, 369–386 (1987)
23. Yin, R.K.: Case Study Research: Design and Methods. Sage, California (2009)
24. Langley, A.: Strategies for theorizing from process data. Academy of Management Review 24, 691–710 (1999)
25. Zafarullah, H.: Adminsitrative reform in Bangladesh: An unfinished agenda. In: Farazmand, A. (ed.) Administrative Reform in Developing Nations, pp. 49–72. Praeger Publishers, Westport (2002)
26. Scott, W.R.: Institutions and Organizations: Ideas and Interest. Sage, Thousand Oaks (2008)
27. Lee, A.S., Baskerville, R.L.: Generalizing generalizability in information systems research. Information Systems Research 14, 221–243 (2003)

A Practical Solution against Corrupted Parties and Coercers in Electronic Voting Protocol over the Network

Thi Ai Thao Nguyen and Tran Khanh Dang

Faculty of Computer Science and Engineering, HCMC University of Technology
268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam
{thaonguyen,khanh}@cse.hcmut.edu.vn

Abstract. In this paper, we introduce a novel electronic voting protocol which is resistant to more powerful corrupted parties and coercers than any previous works. They can be the voting authorities inside the system who can steal voters' information and the content of their votes, or the adversaries outside who try to buy the votes, and force voters to follow their wishes. The worst case is that the adversaries outside collude with all voting authorities inside to destroy the whole system. In previous works, authors suggested many complicated cryptographic techniques for fulfilling all security requirements of electronic voting protocol. However, they cannot avoid the sophisticated inside and outside collusion. Our proposal prevents these threats from happening by the combination of blind signature, dynamic ballots and other techniques. Moreover, the improvement of blind signature scheme together with the elimination of physical assumptions makes the newly proposed protocol faster and more efficient. These enhancements make some progress towards practical security solution for electronic voting system.

Keywords: Electronic voting, blind signature, dynamic ballot, uncoercibility, receipt-freeness.

1 Introduction

Along with the rapid growth of modern technologies, most of the traditional services have been transformed into remote services through internet. Voting service is among them. Remote electronic voting (also called e-Voting) system makes voting more efficient, more convenient, and more attractive. Therefore, many researchers have studied this field and tried to put it into practice as soon as possible. However, that has never been an easy work. It is true that e-voting brings many benefits for not only voters but also voting authorities. Nevertheless, benefits always come along with challenges. The biggest challenge of e-voting relates to security aspects.

In previous works, authors proposed several electronic voting protocols trying to satisfy as many security requirements as possible such as eligibility, uniqueness, privacy, accuracy, fairness, receipt-freeness, uncoercibility, individual verifiability, universal verifiability. However, security leaks cannot be rejected thoroughly in recent

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 11–20, 2013.

electronic voting protocols when voting authorities collude with each other. In the protocol of Cetinkaya et al [5], for example, though the authors announced that their protocol fulfilled the requirement of uncoercibility, once the adversaries corrupted the voter and colluded with the voting authorities taking responsibilities of holing ballots and voter's cast, they could easily found out whether that voter followed their instruction or not. In addition, in voting protocol of Spycher et al [1] and JCJ protocol [6], if the coercer can communicate with the registrars, no longer can voter lie about their credentials. Therefore, the uncoercibility cannot be satisfied. Moreover, in order to satisfy the receipt-freeness, some protocols employed the physical assumptions such as untappable channels that are not suitable for the services through internet.

Most of the previous electronic voting protocols applied three main cryptographic techniques to solve the security problems. Thus, we classify these protocols into three types as: protocols using mix-nets, blind signatures, and homomorphic encryption. The concept of mix-nets was firstly introduced by Chaum in [11]. Since then, there have been some proposed voting protocols such as [7]. However, these protocols met with the big difficulties because of the huge costs of calculations and communications which the mix-net required. Moreover, the final result of voting process is dependent on each linked server in mix-net. If any linked server is corrupted or broken, the final result will be incorrect. So far, no election system based on mix-net has been implemented [13]. Besides mix-net, homomorphic encryption is another way to preserve privacy in electronic voting system. Though homomorphic encryption protocols like [9][14] are more popular than mix-net, they are still inefficient for large scale elections because computational and communicational costs for the proof and verification of vote's validity are quite large. In addition, homomorphic encryption protocols cannot be employed on multi-choices voting forms. As for the blind signature protocols, they also provided anonymity without requiring any complex computational operators or high communicational cost. Until now, there have been many protocols based on blind signature such as [5][8][10][15]. Some of them employed blind signature on concealing the content of votes, others concealed the identifications of voters. Protocol [10], for example, conceals the content of votes; then at the end of voting process, voters had to send the decryption key to the voting authority. This action might break the security if the adversaries conspired with these voting authorities. Therefore, our proposal applies the blind signature technique which is used to hide the real identification of a voter. Besides that, in order to protect the content of votes we apply dynamic ballots along with a recasting mechanism without sacrificing uniqueness to enhance security in the electronic voting protocol and make good the previous protocol 's shortcomings as well.

In this paper, we propose an inside and outside collusion-free electronic voting protocol which guarantees all security requirements. The remarkable contribution is that our proposal is able to defeat the more powerful adversaries which can collude with most of the voting authorities. Another improvement is the enhancement of blind signature scheme that makes our protocol faster and more efficient.

The structure of this paper is organized as follows. In Section 2, we summarize the background knowledge of electronic voting. We describe the details of our proposal protocol in Section 3. Then, in Section 4 security of the protocol is discussed. Finally, the conclusion and future work are presented in Section 5.

2 Background

2.1 Security Requirements

According to [8], the security requirements of electronic voting system are introduced as follows: (1) privacy: no one can know the link between the vote and the voter who casted it; (2) eligibility: only eligible and authorized voters can carry out their voting process; (3) uniqueness: each voter has only one valid vote; (4) accuracy: the content of vote cannot be modified or deleted; (5) fairness: no one, including voting authorities, can get the intermediate result of the voting process before the final result is publicized; (6) receipt-freeness: the voting system should not give voter a receipt which he uses to prove what candidate he voted; (7) uncoercibility: the adversary cannot force any voters to vote for his own intention or to reveal their votes; (8) individual verifiability: every voter is able to check whether their vote is counted correctly or not; (9) universal verifiability: every voter who is interested in tally result can verify it is correctly computed from all the ballots casted by eligible voters or not.

2.2 Cryptography Building Block

Bulletin Boards. In [2], Bulletin board is a communication model which can publish information posted on its body, thus everybody can verify these information. Electronic voting system applies this model to fulfill the requirement of verifiability. In the protocol using bulletin board, voters and voting authorities can post information on the board. Nevertheless, no one can delete or alter these things.

Blind Signature. The concept of blind signature was first introduced by Chaum in 1982. It stemmed from the need of verifying the valid of a document without revealing anything about its content. A simple method to implement the blind signature scheme is to apply the asymmetric cryptosystem RSA. We have some notations: (1) m: the document needs to be signed; (2) d: the private key of authority (signer); (3) (e, N): the public key of authority; (4) s: the signature of m.

The RSA blind signature scheme is implemented as follows:

The owner generates a random number r which satisfies $gcd(r, N) = 1$. He blinds m by the blind factor $r^e \pmod N$. After that, he sends the blinded document $m' = m.\ r^e \pmod N$ to the authority. Upon receiving m', the authority computes a blinded signature s', as illustrated in Eq. (1), then sends it back to the owner.

$$s' \equiv (m')^d \bmod N \equiv (m.\ r^e)^d \bmod N \equiv (m^d.\ r^{ed}) \bmod N \equiv (m^d.\ r) \bmod N \quad (1)$$

According to Eq. (1), the owner easily obtains the signature s, as Eq. (2).

$$s \equiv s' r^{-1} \bmod N \equiv m^d \bmod N \quad (2)$$

Dynamic Ballot. The concept of dynamic ballot was introduced in [5]. This is a mechanism that helps voting protocol fulfill the requirement of fairness. In most of e-voting protocols, authors have used usual ballots in which the order of candidates is

pre-determined. Therefore, when someone gets a voter's casting, they instantly know the actual vote of that voter. Alternatively, the candidate orders in dynamic ballot change randomly for each ballot. Hence, adversaries need the voter's casting as well as the corresponding dynamic ballot in order to obtain the real choice of a voter.

In voting process, each voter can randomly take one of these ballots. He chooses his favorite candidate. Then he casts the order of this candidate in his ballot (not the name of this candidate) to a voting authority and his ballot to another voting authority.

Plaintext Equality Test (PET). The notion of PET was proposed by Jakobsson and Juels [4]. The purpose of PET protocol is to compare two ciphertexts without decrypting. It based on the ElGamal cryptosystem [3].

Let $(r_1, s_1) = (a^{y1}, m_1.a^{x.y1})$ and $(r_2, s_2) = (a^{y2}, m_2.a^{x.y2})$ be ElGamal ciphertexts of two plaintexts m_1 and m_2 respectively. The input I of PET protocol is a quotient of ciphertexts (r_1, s_1) and (r_2, s_2), and output R is a single bit such that $R = 1$ means $m_1 = m_2$, otherwise $R = 0$.

$$I = \left(\frac{r_1}{r_2}, \frac{s_1}{s_2}\right) = \left(a^{y_1 - y_2}, \frac{m_1}{m_2}.a^{x(y_1 - y_2)}\right)$$

According to ElGamal cryptosystem, I is the ciphertext of the plaintext (m_1/m_2). Therefore, if someone who owns the decryption key x, they can obtain the quotient of m_1 and m_2 without gaining any information about the two plaintexts m_1 and m_2.

3 The Proposed Electronic Voting Protocol

3.1 Threats in Electronic Voting Protocol

Vote Buying and Coercion. In a traditional voting system, to ensure a voter not to be coerced or try to sell his ballot to another, voting authorities built some election precincts or kiosk in order to separate voters from coercers and vote buyers. Therefore, they could vote based on their own intentions. When electronic voting system is brought into reality, there are no election precincts or voting kiosks, but voters and their devices which can connect to the internet. Hence, the threats from coercers and vote buyers quickly become the center of attention of the voting system.

Corrupted Registration. Registration is always the first phase of a voting process where voting authorities check voters' eligibilities and give voters the certificates to step into the casting phase. However, in case a voter abstains from voting after registration, the corrupted registrars can take advantages of those certificates to legalize the false votes by casting the extra vote on behalf of the abstaining voters. Sometimes, corrupted registrars can issue false certificates to deceive other voting authorities.

Corrupted Ballot Center. Some protocols have a ballot center as providing voters with ballots. Others, in [5], utilize it for holding the choices that voters made until the casting phase completes. If the ballot center becomes a corrupted party, it can modify the content of the votes or sell them to vote-buyers and coercers who want to check whether the coerced voters cast the candidate they expect. Hence, a feasible electronic voting protocol has to possess the mechanism to protect the system against this threat.

Corrupted Tallier. Tallier takes the responsibility for counting up all the votes to get the final result of voting process. If tallier becomes a corrupted party, it will be able to do that job though the voting process does not come to the end. In this case, it will release the intermediate voting result which has the influence on the psychology of the voters who have not casted the ballots yet. This threat makes the fairness fail.

3.2 The Proposed Electronic Voting Protocol

Before explaining each step in the protocol, we introduce some notations: (1) (e_x, d_x): a public-private key pair of user X; (2) $E_x(m)$: an encryption of *m* with the public key e_x; (3) $D_x(m)$: a decryption/sign of *m* with the private key d_x; (4) H(m): an one way hash function with an input *m*; (5) $E_{PET}(m)$: an encryption of *m* using ElGamal cryptosystem; (6) PET(x, y): a PET function applying PET protocol with two inputs *x, y*.

Registration Phase. In this phase, the blind signature technique is applied to conceal the real identity of a voter through creating an anonymous identity for communicating with other voting authorities. The following paragraphs will show how voters get their anonymous identification from *Privacy of Voter server* (hereafter called *PVer*).

Firstly, the voter sends his real ID to *Registration server* (hereafter called *RS*) to start registration process. Based on the real ID, *RS* checks whether that user is registered or not. If he did this job before, *RS* will terminate his session; otherwise, *RS* will ask *CA* to check the current rules of the voting process in order to find out whether this person can become an eligible voter or not. Then, *RS* creates a certificate and sends it to the voter. This certificate includes: a serial number, a digital stamp, a session key, a signature of *RS*.

Upon receiving the certificate, voter generates his unique identification number:

$$\text{uid} = \text{Hash}(D_V(\text{Digital stamp}))$$

To get the signature of a voting committee on uid, a voter applies the blind signature technique as introduced in Section2.2. He uses a random blind factor to blind uid, and then sends it together with the certificate to *PVer*, which takes the responsibility for preserving privacy of voters. *PVer* saves the serial number in certificate in order to ensure that each certificate asks for the blind signature just one time. After checking the validity of certificate, *PVer* blindly signs the uid, then send the result s' to the voter. He, then, unblinds s' to get the signature s of the voting committee on his uid. Since then, the voter sends uid and corresponding s to other voting authorities for authentication. The detail steps are illustrated in Fig.1.

To avoid man-in-the-middle-attacks, the asymmetric cryptosystem is used at the 1^{st}, 6^{th}, and 8^{th} steps. However, at the 10^{th} step, asymmetric key pairs are not a good choice because they are used only one time for encrypting message, not authenticating. Therefore, the symmetric-key cryptosystem with Tripple DES algorithm is proposed in this blind signature scheme because it has some significant benefits: (1) it does not consume too much computing power so we can shorten encryption time and simplify the period of encryption certificate as well; (2) although symmetric encryption is not as safe as an asymmetric encryption, high level of security still be guaranteed for some reasons that: Triple DES has high complexity, the session key generated randomly by system is long enough to against Brute Force and Dictionary Attack, and the period of using session key is limited in one step with a short time.

Another improvement of this blind signature scheme is that a voter generates list of anonymous identifications including uid, uid_1, and uid_2 instead of just one. The purpose of uid is to communicate with other voting servers; and uid_1 and uid_2 are to ensure the dynamic ballot of voter is not modified by any adversaries.

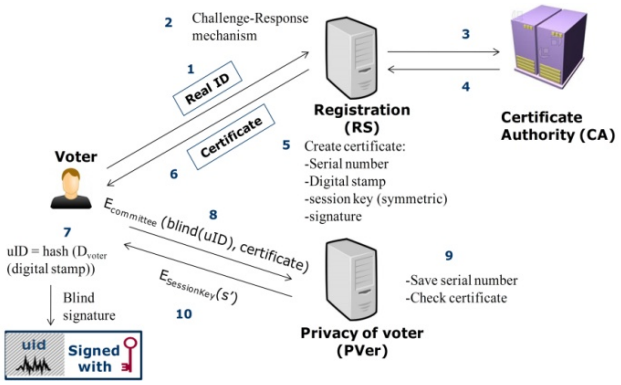


Fig. 1. Registration scheme

Authentication and Casting Phase. To protect privacy of votes from coercers, voting buyers, or sometimes adversaries who stay inside the system, we propose the scheme as shown inFig.2, which applies dynamic ballots, plaintext equivalent test, bulletin boards as introduced in Section 2.2.

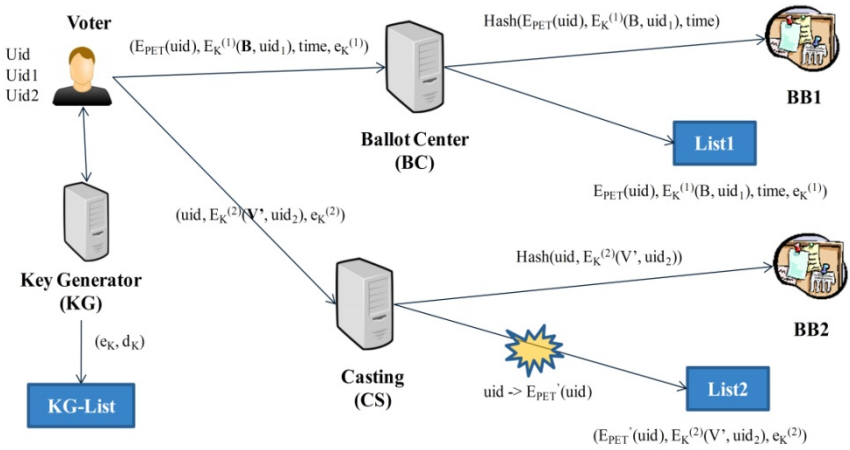


Fig. 2. Casting scheme

In previous works, to avoid coercion as well as vote buying, authors apply a fake identification mechanism to deceive coercers (in JCJ protocol [6]); and others utilize a recasting mechanism without sacrificing uniqueness to conceal voters' final decision [5]. The fake identification requires the condition that at least one voting authority knows the real identification of a voter in order to determine what the real votes are.

Therefore, if this voting authority becomes adversary, the requirements of uncoercibility and vote-buying can be violated. As a result, our proposal uses recasting mechanism to achieve higher level of security.

Firstly, an eligible voter receives the list of candidates from *Ballot Center* (called *BC*). He then, mixes the order of candidates randomly, sends this dynamic ballot B to *BC* and casts his cloaked vote V' by picking the order of the candidate in B he is favor, and sending it to the *Casting server* (called *CS*). To ensure B and V' cannot be modified by others, the voter encrypts them with uid_1 and uid_2 by the couple of public keys $e_k^{(1)}$ and $e_k^{(2)}$ generated by *Key Generator server* (called *KG*). *KG* also saves the private key d_k along with corresponding e_k in KG-List for decrypting in the next phase. After that, voter sends ($E_{PET}(uid)$, $E_k^{(1)}(B, uid_1)$, time, $e_k^{(1)}$) to *BC*, and (uid, $E_k^{(2)}(V', uid_2)$, $e_k^{(2)}$) to *CS* as illustrated in Fig.2. Each message receiving from voter, *CS* checks uid of this voter. If uid is invalid, *CS* discards it; otherwise, it hashes the whole message, and publishes the result on the Bulletin Board BB2 for individual verifiability. It also stores the message into List2 for matching with B in tallying phase. As for *BC*, it does the same things with every message it receives, except authenticating the eligibility of voters.

Voters are allowed to recast. Because the actual vote V of a voter consists of B and V', voters just need to change one of two components to modify the value of V. In this protocol, voters are able to change the orders of candidates in their dynamic ballot B. In order to recast, a voter sends another message (E_{PET}*(uid), $E_k^{(1)}(B^*, uid_1)$, time*, $e_k^{(1)}$) to *BC* in which E_{PET}*(uid), B* and time* are respectively new ElGamal encryption of uid, new dynamic ballot and the time when he sends the message.

Tallying Phase. At the end of casting phase, *PET* server applies PET protocol in Section 2.2 to each $E_{PET}(uid_i)$ in List1. The purpose is to find which pair of encryptions of uid_i and uid_j is equivalent without decryption. After that, *PET* server removes the record holding the earlier time parameter. Concretely, we consider two records R_i and R_j of List1:

$$R_i = (E_{PET}(uid_i), E_{Ki}^{(1)}(B_i, uid_{1i}), time_i, e_{Ki}^{(1)})$$
$$R_j = (E_{PET}(uid_j), E_{Kj}^{(1)}(B_j, uid_{1j}), time_j, e_{Kj}^{(1)})$$

If $PET(E_{PET}(uid_i), E_{PET}(uid_j)) = 1$, and $time_i > time_j$; then the system removes R_j from the system. The purpose of this process is to remove duplicated votes and gain the latest choices of all voters. After that, *PET* server continues to compare each $E_{PET}(uid_i)$ in List2 to each $E_{PET}(uid_j)$ in List1 to find out which B in List1 is corresponding to V' in List2. If there exists a record in List1 which does not match with any record in List2, this record must have come from an invalid voter, so it is discarded at once. The purpose of this process is to remove invalid dynamic ballots B in List1.

After determining pairs of records, KG-List publishes the list of session keys (e_K, d_K) for List1 and List2 to find d_K related to each e_K which is attached to every record in List1 and List2. With the corresponding d_k, $E_K^{(1)}(B, uid_1)$ and $E_K^{(2)}(V', uid_2)$ are decrypted. *Tallying server* (called *TS*) checks the valid of uid_1 and uid_2 to ensure B and V' not to be modified by any parties, then combines the valid values of B and V' to find out the actual vote V of a voter. Finally, *TS* counts up all the actual votes and publishes the result of voting process.

4 Security Analysis

In this section, we provide the security analysis of our proposal and draw the comparisons with the previous typical electronic voting protocols.

Table 1. Comparing the earlier typical protocols with our proposal

Security flaw/requirement	Hasan [12]	Cetinkaya [5]	JCJ [6]	Our protocol
No Vote buying/ Coercion	-	-	√	√
No corrupted *RS*	√	√	-	
No corrupted *BC*	-	-	√	
No corrupted *Tallier*	-	-	√	
No physical assumption	√	√	-	
Privacy	√	√	√	
Eligibility	√	√	-	
Uniqueness	√	√	√	
Uncoercibility	-	-	√	
Receipt-freeness	-	-	√	
Accuracy	-	√	√	
Fairness	-	√	√	
Individual verifiability	-	√	√	
Universal verifiability	-	√	√	

In our protocol, a voter employs the blind signature technique to get the voting authority's signature on his created identity. Therefore, the *RS* and *PVer* do not know anything about the anonymous identity that voters use to authenticate themselves. Hence, if these voting authorities become corrupted, they cannot take advantages of abstention to attack the system. So do the protocols of Hasan [12] and Cetinkaya [5]. In JCJ protocol [6], the *RS* establishes the credentials and passes it to voters through an untappable channel. In the worst case, if the *RS* is a corrupted party, it can give voters fake credentials, and use the valid ones to vote for other candidates. Thus, the corrupted *RS* becomes a security flaw of JCJ protocol. Using physical assumption, i.e. an untappable channel, is another weak point of JCJ protocol in comparison with the previously proposed protocols.

In the voting protocols of Hasan [12] and Cetinkaya [5], though eliminating the abstention attack from corrupted *RS*, these protocols are not stronger enough to defeat sophisticated attacks. The voting protocol of Hasan is quite simple; it has no mechanism to protect the content of votes against being modified. Thus, if *CS* or *TS* collude with attackers, the system will collapse. As a result, the accuracy and fairness properties cannot be guaranteed. In ideal case which every server is trusted, the protocol cannot avoid vote-buying and coercion if voters reveal their anonymous identities to vote-buyer or coercer. As for the protocol of Cetinkaya, it guarantees some security requirements (as illustrated in Table 1). However, the weakness point of this protocol is that the voters are still coerced if the servers holding ballots connive with coercer. In the worst case, voters also able to sell their ballots by providing buyers with their anonymous identities, and then if buyers collude with Ballot Generator, Counter, and

Key Generator, they can find out whether these anonymous identities are attached to the candidate they expect or not. In other words, corrupted *BC* is a security flaw that Cetinkaya has not fixed yet. Our protocol makes good Cetinkaya's protocol shortcomings by encrypting uid using ElGamal cryptosystem before sending it to *BC*. Therefore, when a voter recasts, *BC* itself cannot recognize his uid. Only *Casting* server has responsibility to authenticate the eligibility of uid. However, the recasting process does not take place in *CS*, coercers cannot collect any information from this server.

If *TS* becomes corrupted, our protocol cannot be broken even though *TS* colludes with others voting authorities in protocol. In previous protocol using dynamic ballot, corrupted *TS* just needs to bribe *BC* and *CS* for getting intermediate result. However, in our protocol, B and V' are encrypted with the session key generated by *KG*, so *BC* and *CS* cannot provide the value of B and V' for *TS* without the decrypt key. Even if *KG* is also corrupted, the intermediate result of our protocol is still safe because the uid of voters are encrypted using ElGamal cryptosystem. Attackers have no way to combine B and V' or to remove invalid and duplicated votes. Therefore, corrupted Tallier is no longer a threat for our protocol. However, regarding sophisticated attacks which many voting authorities conspire together, Hasan [12] and Cetinkaya [5] are not strong enough to defeat these kinds of attacks.

According to the blind signature technique, no voting authorities know the link between voter's real ID and his uid and, no one can find out the link between a vote and a voter who casted it. It means that the privacy requirement is guaranteed.

This protocol has multiple layers of protection. For instance, *RS* checks the validity of requesters by CRAM; then, *PVer* check the eligibility of voters by their certificates. Another interesting point of our protocol is that there is a voter's signature d_V in the uid of a voter so the *RS* cannot create a fake uid to cheat other voting authorities without detecting. In brief, our protocol achieves eligibility.

Recasting is allowed in our protocol. If an adversary coerces voters to cast for his intention, the voters can send another vote to replace the previous one. According to the analysis above, this process cannot be discovered by coercers though they connive with many voting authorities. Therefore, the uncoercibility requirement is guaranteed.

Receipt-freeness is also fulfilled when the voters cannot prove their latest casting to vote-buyer. In case that an adversary penetrates into List1 and gets voters' uid through bribing, if the uid is not encrypted, the adversary can easily find out a certain uid does recasting process or not. Consequently, he can threat the voter or discover what the latest casting of voter is. Nevertheless, this assumption has never occurred in our protocol, according to the analysis at the beginning of this section.

The requirement of individual verifiability is guaranteed by applying bulletin boards. *BC* publishes Hash(E_{PET}(uid), $E_K^{(1)}$ (B, uid_1), time) in BB1 and Hash(uid, $E_K^{(2)}$ (V', uid_2)) is published in BB2. Thus, voters just have to hash the necessary information which they have already known, and compare their results to all records in bulletin boards to check whether the system counted his vote correctly.

At the end of election, all voting authorities publish their lists. Any participant or passive observer can check the soundness of final result based on the information on these lists and the bulletin boards as well. Hence, universal verifiability is fulfilled.

5 Conclusion

In this paper, we have proposed an unsusceptible electronic voting protocol to most of sophisticated attacks. The proposed protocol protects the privacy of voters and the content of votes from both inside and outside authorities even though more and more adversaries collude together. Furthermore, the fact no physical assumptions and no complex cryptographic techniques need to be used makes our proposal more practical. In the future, we intend to formalize an electronic voting protocol using process calculi such as pi-calculus for describing concurrent processes and their interactions.

References

1. Spycher, O., Koenig, R., Haenni, R., Schlapfer, M.: A new approach towards coercion-resistant remote e-voting in linear time. In: Financial Cryptography, pp. 182–189 (2012)
2. Araújo, R., Ben Rajeb, N., Robbana, R., Traoré, J., Youssfi, S.: Towards Practical and Secure Coercion-Resistant Electronic Elections. In: Heng, S.-H., Wright, R.N., Goi, B.-M. (eds.) CANS 2010. LNCS, vol. 6467, pp. 278–297. Springer, Heidelberg (2010)
3. Kohel, R.D.: Public key cryptography. In: Cryptography. Book, pp. 67–74 (2010)
4. Meng, B.: A critical review of receipt-freeness and coercion-resistance. Journal Information Technology 8(7), 934–964 (2009)
5. Cetinkaya, O., Doganaksoy, A.: A practical verifiable e-voting protocol for large scale elections over a network. In: 2nd International Conference on Availability, Reliability and Security, pp. 432–442 (2007)
6. Juels, A., Catalano, D., Jakobsson, M.: Coercion-resistant electronic elections. In: ACM Workshop on Privacy in the Electronic Society, pp. 61–70 (2005)
7. Camenisch, J., Lysyanskaya, A.: A formal treatment of onion routing. In: 25th Annual International Cryptology Conference, pp. 169–187 (2005)
8. Liaw, H.T.: A secure electronic voting protocol for general elections. Journal Computers and Security 23, 107–119 (2004)
9. Baudron, O., Fouque, P.A., Pointcheval, D., Poupard, G., Stern, J.: Practical Multi-Candidate Election System. In: 20th ACM Symposium on Principles of Distributed Computing, pp. 274–283 (2001)
10. Fujioka, A., Okamoto, T., Ohta, K.: A Practical Secret Voting Scheme for Large Scale Elections. In: Zheng, Y., Seberry, J. (eds.) AUSCRYPT 1992. LNCS, vol. 718, Springer, Heidelberg (1993)
11. Chaum, D.: Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms. Journal Communications of the ACM 24(2), 84–88 (1981)
12. Hasan, M.S.: E-Voting Scheme over Internet. In: Conference on Business and Information (2008)
13. Brumester, M., Magkos, E.: Towards secure and practical e-Elections in the new era. Secure Electronic Voting 7, 63–76 (2003)
14. Acquisti, A.: Receipt-free homomorphic elections and write-in voter verified ballots. ISRI Technical Report CMU-ISRI-04-116, Carnegie Mellon University, PA, USA (2004), http://eprint.iacr.org/2004/105.pdf
15. Juang, W.S., Lei, C.L., Liaw, H.T.: A verifiable multi-authority secret election allowing abstention from voting. The Computer Journal 45, 672–682 (2002)

Early-Detection System for Cross-Language (Translated) Plagiarism

Khabib Mustofa and Yosua Albert Sir

Department of Computer Science and Electronics,
Universitas Gadjah Mada
khabib@ugm.ac.id, yosuasir@gmail.com

Abstract. The implementation of internet applications has already crossed the language border. It has, for sure, brought lots of advantages, but to some extent has also introduced some side-effect. One of the negative effects of using these applications is cross-languages plagiarism, which is also known as *translated plagiarism*.

In academic institutions, translated plagiarism can be found in various cases, such as: final project, theses, papers, and so forth. In this paper, a model for web-based early detection system for translated plagiarism is proposed and a prototype is developed. The system works by translating the input document (written in Bahasa Indonesian) into English using Google Translate API components, and then search for documents on the World Wide Web repository which have similar contents to the translated document. If found, the system downloads these documents and then do some preprocessing steps such as: removing punctuations, numbers, stop words, repeated words, lemmatization of words, and the final process is to compare the content of both documents using the modified sentence-based detection algorithm (SBDA). The results show that the proposed method has smaller error rate leading to conclusion that it has better accuracy.

Keywords: translated plagiarism, sentence-based detection algorithm (SBDA), modified-SDBA,Google API.

1 Introduction

Since it was first introduced, internet has brought several changes in human life, not to mention in academic environment. The existence of search engines makes students and teachers easy in finding materials for enhancing their knowledge, but in other point of view, it also facilitates any attemps of academic misconduct such as plagiarism.

Generally, academic misconduct and plagiarism may happen within several types:

1. *copy-paste* : copying part or the whole content of document
2. *paraphrasing* : changing the grammar, changing the order of constructing sentences, or rewriting the documents' content using synonim

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 21–30, 2013.

3. *translated plagiarism*: translating part or the whole content of document from some language to some other language

The above approaches can fall into plagiarism if the writer does not provide correct citation or without mentioning the source document references ([1]). Plagiarism type (1) or (2) can easily be done, and type (3) is little bit more complicated as the writer still needs to translate the source into different language. According to [1], several tools already exist to suspect or detect plagiarism type (1) or (2), such as : *Turnitin, MyDropBox, Essay Verification (EVE2), WcopyFind, CopyCatch, Urkund*, and *Docoloc*, while plagiarism of type (3) was discussed in ([2]), eventhough without revealing the quantitative accuracy.

This paper will discuss a model and establishment of a system for early detecting translated plagiarism. The system works under the following constraints:

- source document is written in Bahasa Indonesia.
- the system is not a "silver bullet" to translated plagiarism. The output of the system is not an absolute judgement whether a palgiarism does exist.
- the algorithm used in comparing documents is based on *sentence based detection algorithm*
- to enhance the accuracy of detection process, the algorithm is slightly modified by incorporating synonyms of the words constructing sentences. This approach is carried out during stemming and lemmatization process.

2 Problem Formulation

Suppose there exists a suspect document D_q, written in Bahasa Indonesia. On the other hand, there is also a set of vast amount of documents (reference documents), written in other language, Ω, available on the web repository. In this case, for simplicity, we will assume that all documents, $\forall d_i \in \Omega$, are written in English. When there exist statements in or part of a document D_q, whose translation is similar in meaning to statements from some documents in Ω, *how can we find and identify such statements, either in suspected document and also in reference documents*?

Using sentence-based detection algorithm, the target can be obtained by computing similarities which result from comparing each statements in suspected document, $\forall s_q \in D_q$, with all statements in all reference documents, $\forall s_r \in D_r : \forall D_r \in \Omega$. Theoretically, the above process is feasible, but in practice, special treatment should be incorporated as the size or dimension of Ω is big enough and also both the D_q and D_r are in different languages.

The following questions give us guidance in understanding the approach to be discussed further in this paper:

1. How is the system architecture to achieve the goal of detecting translated plagiarism?
2. How to translate D_q (in Bahasa Indonesia) into D_q^* (in English) , where later the statements in D_q^* will be compared with statements in D_r ?

3. How to reduce the size of Ω, for example by constructing Θ, where $\Theta \subset \Omega$, such that $|\Theta| \lll |\Omega|$ and $\forall D_x \in \Theta$, D_x is a document which has similar (not necessarily all) content with D_q^*
4. How to calculate the similarity between $\forall s_q \in D_q^*$ and $\forall s_r \in D_x : \forall D_x \in \Theta$?

Figure 1 shows the architecture of the system as the realization to answer the above questions.

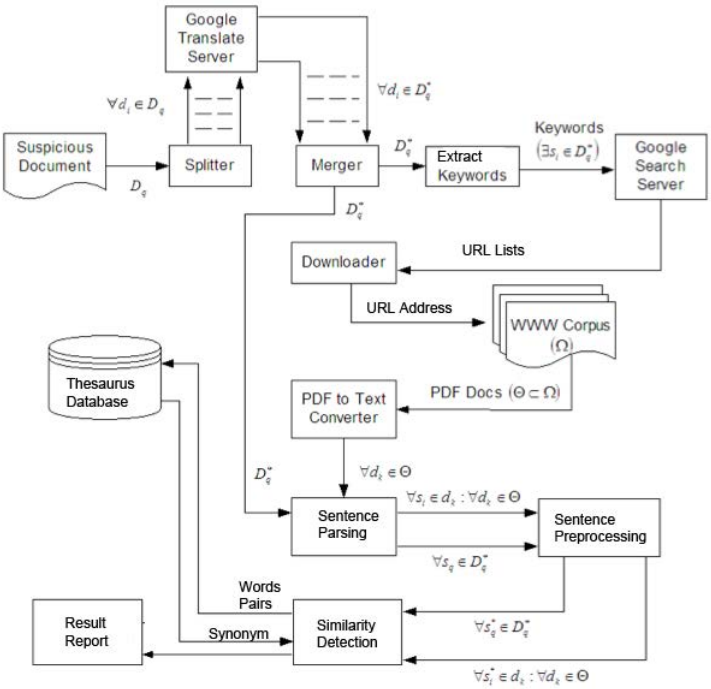


Fig. 1. Architecture of the Early-Detection System for Translated Plagiarism

3 Related Works

In general, the methods for detecting plagiarism on text documents can be categorized into two ways: *document fingerprint comparison* and *full-text comparison* [3]. In the first approach, contents of the input documents (both suspected and reference documents) are converted into a more compact form, based on some predefined method, and then the detection process is carried out by comparing the fingerprints without comparing the whole contents.

Kent and Salim ([2]) explored the first approach by forming *4-grams document fingerprinting* and then comparing the fingerprints using *Jaccard Distance*. The result showed that the approach incurred disadvantages:

1. vulnerable against changes in words order or sentence order. Changes in words or sentence order, even just a small changes, may result in significant changes in fingerprints.
2. unable to detect small change that may be aded by plagiarist, such as word insertion or deletion

White and Joy ([3]) implemented the second approach by using *sentence-based detection algorithm*. In this method, the detection process on plagiarism between two documents is approached by calculating similarities of all pairs of sentences constructing the documents.

This paper will discuss an approach which extends the second approach by adding feature of incorporating synonym to enhance the capability of finding sentences eventhough the sentences have been modified by changing some words with their synonyms. This means, compared to the method implemented by Kent and Salim ([2]), the proposed approach will overcome the disadvantages of using fingerprinting. While compared to White and Joy ([3]) which implements *sentence-based detection algorithm*, the proposed approach will enhance the method by adding capability of invetigating synonims of words appearing in the sentences.

4 Methodology

4.1 Documents Similarity Measures

A document can be viewed as a series of tokens which may come in form of letters, words or sentences. If it is assumed that there is a parser capable of parsing the document contents, ignoring puctuation marks, formatting commands and capitalization, then the output of such parser is a canonical sequence of tokens [4]. In his paper [4], Broder introduced two metrics for measuring the similairty between documents A and B : *resemblance* ($r(A, B)$) and *containment* ($c(A, B)$), expressed as follows:

$$c(A, B) = \frac{|d(A) \cap d(B)|}{|d(A)|} \tag{1}$$

$$c(B, A) = \frac{|d(A) \cap d(B)|}{|d(B)|} \tag{2}$$

$$r(A, B) = \frac{|d(A) \cap d(B)|}{|d(A) \cup d(B)|} \tag{3}$$

where

- $d(X)$ symbolizes set of token in document X
- $r(A, B)$ has value $x \in \Re, x \in [0, 1]$. If $r(A, B) = 1$ then $d(A) = d(B)$
- $c(A, B)$ has value $x \in \Re, x \in [0, 1]$. If $c(A, B) = 1$ then $d(A) \subseteq d(B)$. Two documents A and B are said to be identical if and only if the set of all tokens in A is subset of the set of all tokens in B and vice versa.

– Assumed that the canonical sequence of token is in form of a sentence, $|d(x)|$ indicates the number of sentences in document X (length/size of the set), and $|d(A) \cap d(B)|$ can be considered as common sentences found in document A and B.

As an illustration, given that

Document A : *a rose is a rose is a rose*
Document B : *a rose is a flower which is a rose*

Both sentences can be pre-processed as shown in table 1, and the value of $c(A, B) = 87.5\%$), $c(B, A) = 77.78\%$ and $r(A, B) = 70\%$

Table 1. Example of documenst preprocessing to obtain *resemblance* and *containment*

A	B	$d(A) \cap d(B)$	$d(A) \cup d(B)$
a:3	a:3	a:3	a:3
rose:3	rose:2	rose:2	rose:3
is:2	is:2	is:2	is:2
	flower:1		flower:1
	which:1		which:1

4.2 Sentence-Based Detection Algorithm

In order to apply this method, in which the comparison of all pairs of sentences from both document should be done, three steps must be performed:

1. Documents Preprocessing. This step includes: decapitalization, removing stop words, removing duplicate words
2. Computing Sentences Similarity. Assume that A and B are documents to be compared, comprised of sentences. we use the following symbols to shorten the upcoming discussion.
   – $d(X)$: a set of sentences in document X
   – s_i^x : i^{th} sentence of document X
   – $d(s_i^x)$: set of words comprising s_i^x, can also be denoted as $s_i^x = \{w_1^x, w_2^x, \ldots, w_n^x\}$

   The measure of similarity can be computed from the following equations:

$$common_{words} = \begin{cases} 1 & \text{if } w_i^a \text{ and } w_j^b \text{ is identical} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$SimScore(A, B) = \left(|common_{words}| \times \frac{|d(A)| + |d(B)|}{2 \times |d(A)| \times |d(B)|} \right) \times 100\% \tag{5}$$

$$SimSent(A, B) = \begin{cases} SimScore(A, B) & \text{, if } SimScore(A, B) \geq SimTh \\ & \text{OR } |common_{words}| \geq ComTh \\ 0 & \text{, otherwise} \end{cases} \tag{6}$$

   where $ComTh$ represents *Common Threshold* and $SimTh$ indicates *Similarity Threshold*

3. Computing Documents Similarity. Whenever the whole pairs of sentences from document A and B have been examined and sentences similarity scores have been obtained, the document similarity score can be calculated by summing all sentences similarity scores.

4.3 Modified Sentence-Based Detection Algorithm

In the previous section, from equation (4), (5) and (6), it is clear that the existence of common words clearly contributes to the similarity scores. Suppose that within the two sentences to be compared there exist words form the first sentence having similar meaning with some words in the second sentence. *How if we treat those pairs of words also as common words (with different weight of commonness)*?

Assuming the above question works, we can derive the proof that changing some words in the first sentence into their synonyms will affect the similarity measures as long as the synonyms are in the second sentence.

Suppose the first sentence is the i^{th} sentence of document A, denoted by $s_i^a = \{w_1^a, w_2^a, w_3^a, \ldots, w_n^a\}$, and the second sentence is the j^{th} sentence of document B, denoted by $s_j^b = \{w_1^b, w_2^b, w_3^b, \ldots, w_m^b\}$. Then:

1. $common_{OLD}(s_i^a, s_j^b) = s_i^a \cap s_j^b$
2. $Syn(w)$ =synonym of w
3. $Diff(s_i^a, s_j^b) = s_i^a - s_j^b$
4. $Diff(s_j^b, s_i^a) = s_j^b - s_i^a$
5. $SynWORD(s_i^a, s_j^b) = \{w_k | w_k \in Diff(s_i^a, s_j^b) \wedge Syn(w_k) \in Diff(s_j^b, s_i^a)\}$
6. $|common_{NEW}| = |common_{OLD}| + |SynWORD(s_i^a, s_j^b)| \times 0.5$, as synonyms are considered common words with different weight of commonness (in this case 0.5)

Hence

$$(SynWORD(s_i^a, s_j^b) \neq \{\}) \Longrightarrow (|common_{NEW}| > |common_{OLD}|)$$

Reformulation of equation (4) and (5) by incorporating synonyms yields slightly different forms:

$$common_{words} = \begin{cases} 1 & \text{if } w_i^a \text{ and } w_j^b \text{ is identical} \\ 0.5 & \text{if } w_i^a \text{ is synonym of } w_j^b \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

$$SimScore(A, B) = \left(\frac{2 \times |common_{words}|}{|d(A)| + |d(B)|} \right) \times 100\% \tag{8}$$

In standard Sentence-Based Detection Algorithm, $SimScore(A, B)$ is calculated using asymmetrical average method (equation 5), while in modified algorithm it is expressed using *Dice* method (equation 8).

The measure or score of documents similarity is then defined as combination of two asymmetric similarity scores [5]:

$$Sim(A,B) = \left\langle \frac{|d(A) \cap d(B)|}{|d(A)|}, \frac{|d(A) \cap d(B)|}{|d(B)|} \right\rangle \tag{9}$$

The same approach is also used in [6] and [7]. By using this form, two-way similarity score can be obtained and gives better picture of the relationship between the two document. Suppose we have two documents A and B, and $|d(A)| = 120$ sentences, $|d(B)| = 160$ sentences and $|d(A) \cap d(B)| = 80$ sentences, then the similarity score of document A and B, $Sim(A,B) = \langle 0.667, 0500 \rangle$ which can be interpreted that two third of sentences in A can also be found in B and half of sentences in B can be found in A.

4.4 Illustration of Algorithm Usage

In this section, we will look on how both methods (standard Sentence-Based Detection and Modified Sentence-Based Detection) are applied to the same dataset and see the differencces of their usages.

Let us assume that document A and B have the following contents:

> A : Face detection is one of the crucial early stages of face recognition systems are used in biometric identification.
> B : Face detection is one of the most important preprocessing step in face recognition systems used in biometric identification.

and assign *similarity threshold* 80, *common threshold* 6 and *stop words* {the, is, on, in, are}. After pre-processing step (decapitalization, removing stop words, removing duplicate words), we have:

1. $SentenceObject_A = \{face,\ detection,\ one,\ crucial,\ early,\ stages,\ recognition,\ systems,\ used,\ biometric,\ identification\}$
2. $SentenceObject_B = \{face,\ detection,\ one,\ most,\ important,\ preprocessing,\ step,\ recognition,\ systems,\ used,\ biometric,\ identification\}$
3. $common_{words}(A,A) = \{face,\ detection,\ one,\ crucial,\ early,\ stages,\ recognition,\ systems,\ used,\ biometric,\ identification\ \}$
4. $common_{words}(A,B) = \{face,\ detection,\ one,\ recognition,\ systems,\ used,\ biometric,\ identification\}$
5. $Diff(SentenceObject_A, SentenceObject_B) = \{crucial,\ early,\ stage\}$
6. $Diff(SentenceObject_B, SentenceObject_A) = \{most,\ important,\ preprocessing,\ step\}$.

   By pairing each words in $Diff(SentenceObject_A, SentenceObject_B)$ with each words in $Diff(SentenceObject_B, SentenceObject_A)$ and consulting with translation service (such as Google Translate), additional information about synonym is obtained, $SynWord(A,B) = \{important, step\}$ (as "important" is synonym of "crucial" and "step" is synonym of "stage").

Calculating Similarity Using Sentence-Based Detection Algorithm. Based on the summary of preprocessing result above, the following scores can easily be obtained:

1. $SimScore(A, A) = \left(11 \times \frac{11+11}{2\times11\times11}\right) \times 100 = 100$ and $SimSent(A, A) = 100$
2. $SimScore(A, B) = \left(8 \times \frac{11+12}{2\times11\times12}\right) \times 100 = 69.70$ and $SimSent(A, B) = 69.70$
3. Similarity between A (as reference) and B is $Sim(A, B) = \frac{69.70}{100} \times 100\% = 69.70\%$

Calculating Similarity Using *Modified* Sentence-Based Detection Algorithm (SBDA). Based on the summary of preprocessing result above, the following scores can easily be obtained:

1. $|common_{NEW}| = |common_{OLD}| + |SynWord(A, B)| \times (0.5) = |8| + |2| \times (0.5) = 9$
2. Based on Eq . 8, $SimScore(A, B) = \left(\frac{2\times9}{11+12}\right) \times 100 = 78.26$
3. Similarity between A and B is $Sim(A, B) = \langle 81.82\% : 75\% \rangle$. This score can be interpreted as: 81.82% of sentences in document A can be found in document B, and 75% of sentences in document B can be found in A.

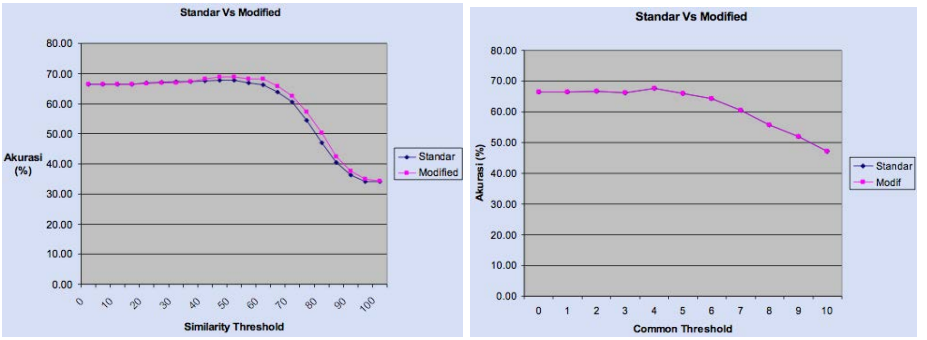
5 Implementation, Testing and Results

Modules Implementation. The system is built by extending and reusing some existing tools. Based on the architecture given at Figure 1, the following are modules developed and tested against some libraries:

1. **Translation Module.** As Google Service restricts the length of translated text, this module first splits long text into possibly several chunks of size maximum 4KB. Each chunk is sent to Google as a request and then the result is combined again to construct the whole document translation.
2. **Document-Searching Module.** Searching of documents on the web repository using search engine requires keywords. This module uses *Named-Entity Recognition (NER)* from Standford, available at `http://nlp.stanford.edu/software/stanford-ner-2009-01-16.tgz`. The tools will search, identify and extract entities of type *person, location* and *organization*. Those types are unique, difficult to plagiarize and suitable to be taken as keywords. This module will perform keyword extraction, look for documents in web repository (based on keywords extracted) and, then download the found PDF documents.
3. **Text Extraction from PDF Documents.** The contents of PDF documents just downloaded are then extracted using existing tools *xpdf version 3.2*.
4. **Content Preprocessing.** This module is responsible for: *decapitalization*, eliminating stop words and punctuation symbols, eliminating repitition of words and lemmatization.

Testing and Results. For the sake of testing, all reference datasets (*corpus*) used in this research are deterministic, in the sense that the similarity degrees of the datasets have been apriori known.

1. **Unit Testing.** Reference Dataset used in this unit testing is taken from *Microsoft Research Paraphrase Corpus*, available at `http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042`. This dataset consists of 1725 pairs of sentences extracted from thousands documents, having been justified by human annotator to classify whether the pair of sentences is a paraphrase or not, resulting 1147 pairs are identified as paraphrases and 578 are not paraphrases. Based on the dataset, by adjusting several values of *similarity threshold* or *common threshold*, as depicrted in Fig. 2, the optimal value for similarity threshold is 50%, and common threshold is 4.



(a) Accuracy vs Similarity Threshold (b) Accuracy vs Common Threshold

Fig. 2. Accuration versus *Similarity Threshold* or *Common Threshold*

Table 2. Average Asymmetric Similarity and Error Rate between Modified and Standard Method

Doc ID	Average Asymmetric Similarity			Error Rate	
	Modified	Standard	Actual	Modified	Standard
1	56.82	42.66	68.18	129.16	651.53
2	61.88	44.89	66.31	19.58	458.60
3	51.24	31.50	65.22	195.30	1136.70
4	51.23	17.45	70.93	388.29	2860.65
5	48.04	29.15	70.22	491.73	1686.33
6	56.97	27.34	68.37	129.85	1683.05
7	60.58	49.09	80.77	407.84	1003.94
8	53.35	31.82	69.77	269.45	1439.82
9	63.20	60.24	72.22	81.45	143.64
10	46.46	47.51	71.80	641.86	590.00
Root Mean Square Error (RMSE)				16.60	34.14

2. **Integration Testing.** For the integration and functional testing, ten documents from `http://pps.unnes.ac.id` are taken as samples. These samples document are processed, and then their similarity are calculated. Table 2 shows the result of similarity test, revealing the outperformance of the proposed method (modified SBDA) against the standard method.

6 Conclusion

Plagiarism is a serious misconduct in academic environment, thus it must be anticipated. The advancement in internet technology and services have been enabling users to more easily conduct plagiarism, but on the other hand, such condition also provides environment to easier check whether any attemp to plagiarism has happened. This paper has shown an approach to early detection of translated plagiarism. The prototype was developed by integrating online services, online repository and implementing modified sentence based detection algorithm. From the result, the following can be concluded:

1. The *modified SBDA* show higher accuracy compared to standard SBDA. This is indicated by smaller value of *error-rate*
2. The modification of standard SBDA is carried out by incorporating synonym conversion. Converting words into their synonyms will *increase the count of common words* between the documents compared, contributing to the better accuracy in document similarity measurement.

References

1. Maurer, H., Kappe, F., Zaka, B.: Plagiarism - a survey. Journal of Universal Computer Science 12(8), 1050–1084 (2006)
2. Kent, C.K., Salim, N.: Web based cross language plagiarism detection. Journal of Computing 1(1) (2009)
3. White, D.R., Joy, M.S.: Sentence-based natural language plagiarism detection. Journal on Educational Resources in Computing 4(4) (2004)
4. Broder, A.Z.: On the resemblance and containment of documents. In: Compression and Complexity of Sequences, SEQUENCES 1997, pp. 21–29. IEEE Computer Society (1997)
5. Monostori, K., Finkel, R., Zaslavsky, A., Hodász, G., Pataki, M.: Comparison of Overlap Detection Techniques. In: Sloot, P.M.A., Tan, C.J.K., Dongarra, J., Hoekstra, A.G. (eds.) ICCS-ComputSci 2002, Part I. LNCS, vol. 2329, pp. 51–60. Springer, Heidelberg (2002)
6. Yerra, R.: Detecting similar html documents using a sentence-based copy detection approach. Master's thesis, Brigham Young University (2005)
7. Smith, R.D.: Copy detection systems for digital documents. Master's thesis, Brigham Young University (1999)

TransWiki: Supporting Translation Teaching

Robert P. Biuk-Aghai[1] and Hari Venkatesan[2]

[1] Department of Computer and Information Science
Faculty of Science and Technology, University of Macau
robertb@umac.mo
[2] Department of English
Faculty of Social Sciences and Humanities, University of Macau
hariv@umac.mo

Abstract. Web-based learning systems have become common in recent years and wikis, websites whose pages anyone can edit, have enabled online collaborative text production. When applied to education, wikis have the potential to facilitate collaborative learning. We have developed a customized wiki system which we have used at our university in teaching translation in collaborative student groups. We report on the design and implementation of our wiki system and an evaluation of its use.

Keywords: wiki, translation, collaborative learning.

1 Introduction

Computer Supported Collaborative Learning (CSCL) became a reality with advances in technology and notably the advent of the Internet. Computer Aided Instruction, the precursor to CSCL, brought about automation of data dissemination, test-taking, language drills etc. that radically changed the classroom. As technology improved around the mid-1990s, computers and the emerging internet began to be seen as potential tools for creating new learning environments [1], moving away from the mostly uni-directional lecture format of teaching. In 1995 Ward Cunningham created the first wiki [2], a web site whose pages could be read and modified by anyone with access to the internet. Wikis could act as repositories for storage and dissemination of information and the collaborative production of assignments, projects, essays etc. We describe one implementation of a wiki that aims to provide a collaborative learning environment for translation.

The following section provides an overview of related work on using wikis in education. Section 3 then introduces our wiki design and implementation, and Sect. 4 presents an evaluation of the use of our wiki system at our university. Finally, Sect. 5 makes conclusions.

2 Related Work

In the field of education, the past few years have witnessed increased adoption of wikis [3]. Generally they are being used in two principal ways: as knowledge repositories for instructors making lecture material and course information

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 31–40, 2013.

available [4]; and for the collaborative production of content by students, such as assignments, projects, essays and other assessment material [5,6]. Examples of the diverse uses which wikis were put to include: collaborative construction of encyclopedia entries by upper secondary school students [6] and university students [5]; collaborative creation of course content [7] or a shared artefact [8] by university students; project management in a project-oriented computer science course [9]; project-oriented product development in an online graduate course [10]; essay writing in a university ESL course [11], and for both short and semester-long assignments in a graduate course [12]; and for developing knowledge management processes [13]. Wikis have been found to be effective in supporting learning [14,15,16], although the success of wikis may depend on assessment and grades as a form of reward for user contributions [17].

The wiki concept as developed by Cunningham is an open one, where anyone can access and modify any page. This is for example largely the way that the MediaWiki system (www.mediawiki.org) underlying the Wikipedia site works although there are facilities for protecting pages from editing. For educational purposes some degree of access control is usually necessary, e.g. to protect pages that should only be editable by instructors but not by students. Moreover, in order to allow identification of contributions and to prevent changes by outsiders, anonymous editing may need to be disabled and students may be required to login [18]. It has also been observed that a single tool integrating all required functions for communication, project management and authoring is preferable to a set of separate tools [10]. Basic wiki technology has thus been extended with several different functions specifically for use in education: protecting/locking pages, creating read-only snapshots of an entire wiki site, and others [7,19].

3 TransWiki

Our translation wiki system, named TransWiki, is based on the open source MediaWiki system which is the wiki engine of Wikipedia and many other popular wiki sites. MediaWiki is designed for a very open editorial process in which by default every user has read and edit privileges to every page. This can, however, be restricted by locking pages, which in the case of Wikipedia is done whenever edit wars break out on hotly contested pages dealing with controversial topics.

Whereas this open design is suitable for wikis with egalitarian contributors, it is not very suitable for use in education where there are clearly distinct roles of instructor and student. Therefore we have tailored MediaWiki for the use in education through a collection of extension packages that can be added onto an existing MediaWiki installation. An extension essentially provides a plugin that extends the functionality of the MediaWiki system. We have made our extensions available as an open source package named UMEduWiki[1]. The system structure of TransWiki is shown in Fig. 1. On the server side, the base MediaWiki system is extended by our UMEduWiki extension package, and each of these has its own database (lower portion of Fig. 1, server and database layers). On

[1] `http://umeduwiki.sourceforge.net/`

the client side a unified TransWiki user interface is presented, which consists of both MediaWiki and UMEduWiki portions, and communicates with each of these server components (upper portion of Fig. 1, client layer).

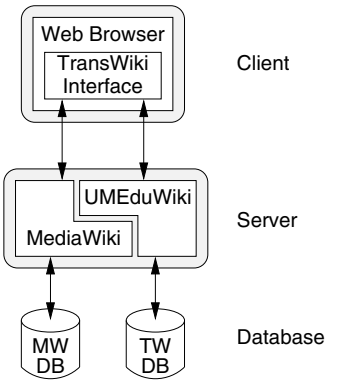


Fig. 1. TransWiki system structure

The base MediaWiki system provides basic wiki functionality: reading and editing of wiki pages, asynchronous online discussion through so-called discussion pages which are attached to wiki article pages, searching, viewing the revision history for any page, as well as several less frequently used wiki editing functions that are normally used by wiki administrators such as locking pages, moving pages, blocking users, and more. TransWiki extends this basic functionality of MediaWiki through following additional functions:

- Group-based access control
- Access log
- Forum-style discussions
- Embedded audio playback

We briefly describe each of these below.

3.1 Access Control

Many wikis adopt a "free for all" access model where (almost) every user can perform (almost) every operation. This is not well suited for use in education where instructors and students have distinct roles and need to perform different operations. For example, an instructor may wish to arrange students in separate groups and assign them to do group work. During the process of the group work, the instructor may want to ensure that no group can view the outputs of any other groups' work. However, after the completion of the assigned work the instructor may wish to lock all contributions to prevent further editing (to perform grading of the assignment), and may decide to simultaneously open read

access to all groups' pages to each other to let them compare their own outcomes with those of their peers. This is the model we have adopted in our translation courses, and it requires a more fine-grained and sophisticated access control than MediaWiki provides.

Our access control component uses a group-based access control model. An instructor or other suitably privileged user can create groups, such as instructor and student groups, or multiple student groups for when a class is divided into several groups. Students are then assigned to these groups. Along with each group we also create a MediaWiki *namespace*, which allows related pages to be grouped together and be collectively controlled. Students can create pages that belong to their namespace, and instructors can then control access to these pages. Access can be either *read* or *edit*. When neither of these is defined then this corresponds to no access rights at all. Access is granted for all the pages of a specified group. For example, an instructor may initially allow groups 1 and 2 only read and edit access to their own pages, then after completing an assignment remove the edit access and at the same time assign read access to each other's pages to share each group's outcome with each other. This kind of access control model may be called *medium-grained*, lying between the coarse-grained access control of the standard MediaWiki system in which all users are divided into a few roles (user, administrator, sysop etc.) with access permissions having a site-wide scope, and fine-grained access control in other systems where the specific permissions for each user can be set differently on an object-by-object basis. In our experience, this medium-grained access control model combines simplicity with effectiveness and is "fine grained enough" for our purposes. Figure 2 shows the user interface of the privilege management part of our access control panel. The user can select one or more groups on the left (G11, G12, G13, Teacher) to whom to assign the access privilege, select the type of access (read and/or edit), and select to which groups' pages the access should be granted, shown on the right (G11, G12, G13, Teacher).

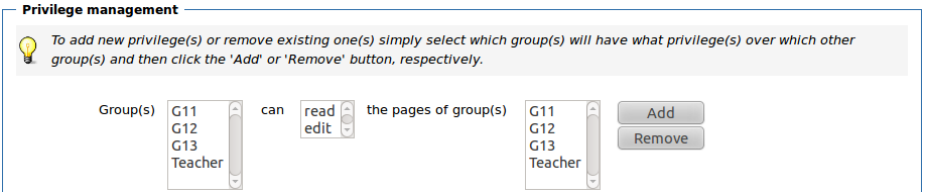


Fig. 2. TransWiki privilege management

3.2 Access Log

Wiki engines such as MediaWiki usually record when a user edits a page, including the version, date and time of the edit, a user-provided summary, and some other information. However, MediaWiki does not record when a user views a wiki page. To assess the involvement of students in online activities instructors

usually wish to know about both reading and editing of wiki pages by students. Therefore we have developed an access log extension as a MediaWiki special page. This special page, which is only visible to instructors, shows who read or edited a page, and when. It provides options to filter the list by user, action, and/or namespace, and limit the number of days to show. It also provides an option to purge old log entries. The Access Log is effective in showing instructors the participation of students in the wiki.

3.3 Discussion Forum

Wiki discussion pages in MediaWiki are simply unstructured wiki pages that can be edited by any user, just as regular wiki pages. This is simple but does not always result in a readable record of the discussion as it lacks threading and sorting of discussion statements. Our earliest attempt to remedy this situation was a customized discussion page in which the page was not editable in the usual way but users could only append new posts to the end of the page. We also automatically added a signature line after each post including user name and a timestamp. This was an improvement over the default discussion page as it provided a chronologically sorted record of discussion posts, but all posts were in a single consecutive sequence without any separate threads, making it difficult to keep track of a discussion on a specific topic. Subsequently we developed an entirely new discussion function, with its own separate database tables. This discussion function supports two types of discussion statements: posts and replies. A post effectively starts a new thread, whereas a reply is a statement made in reply to a given post. This limits our discussion forum to just two levels, unlike other discussion forums that allow deeply nested discussions. Our purpose in doing so was to keep the discussions simple while supporting distinct threads. Moreover, we designed the user interface of our discussion forum to closely resemble that of Facebook posts and comments (see Fig. 3). An “agree” link is provided on each post and reply statement, analogous to the “like” link in Facebook. Students are encouraged to click it to indicate agreement with the statement made. This has the effect that one can quickly gauge which proposed translations find widespread acceptance in the group and which not. Clicking on “agree” is also more subtle than directly expressing agreement or disagreement through written statements, which many of our students avoid in order not to offend their fellow students. A lack of “agree’s” thus implies disagreement, but without explicitly stating so. As practically all our students are active Facebook users, having a Facebook-like appearance achieves instant familiarity with our discussion function.

In online collaborative translation, discussions play an important role in helping move the collective translation work forward by discussing the suitable way to translate a given text. In teaching translation, instructors need ways to assess not only the outcome of the translation, i.e. the target text, but also the process which led up to this outcome, which is recorded in the online discussion. To facilitate assessment of discussions, we have extended the instructor’s view of the discussion forum with two additional functions: a rating function and a tagging

Fig. 3. TransWiki discussion

function. Ratings allow the instructor to indicate how significant a contribution the given discussion post has made and are on a 5-point Likert scale (-2, -1, 0, 1, 2). Tags indicate the type of contribution made, and can be one of the following nine options: Informative, Argumentative, Evaluative, Elicitative, Responsive, Confirmation, Directive, Off task, or Off task technical. Rating and tagging together thus capture both a quantitative as well as a qualitative assessment of each discussion statement.

3.4 Audio Player

Our translation teaching includes the teaching of simultaneous interpretation. To integrate the teaching of interpretation into the wiki, we upload mp3 audio files of spoken interpretation exercises to our wiki. A MediaWiki extension that provides a public-domain mp3 player is seamlessly embedded directly in a wiki page to play back these audio files.

4 Evaluation

We deployed the TransWiki system at our university in the year 2007 and have used it since in two of our undergraduate translation classes which teach Chinese-English translation. Classes were divided into groups of 4–5 students each. These were composed of students with maximum variety of backgrounds, i.e. local students vs. those from mainland China, students from Chinese medium high schools vs. those from English medium high schools, etc. The underlying assumption in this modus operandi is that collaboration among students in producing the translation will result in interaction, peer review and discussion that would help students better understand the subject. As Tudge observes "...Research based on this model has indicated that social interaction between peers who bring different perspectives to bear upon a problem is a highly effective means

of inducing cognitive development" [20]. Furthermore, the social constructivist perspective on learning that we embrace in the teaching of translation holds that it is "by communicating and negotiating with peers and more experienced (and thus more knowledgeable) others, we acquire a feel for correctness, appropriateness and accuracy, a feel that is grounded in our social experiences..." [21]. Thus discussion and collaborative construction of translations are core to the development of translation skills.

Students were given assignments of translating a source text (in either English or Chinese) into the target language (i.e. from English to Chinese, or from Chinese to English). The discussion among group members was to be conducted in the assignment's discussion page inside TransWiki, and at the end when group consensus on the translation had been reached, the resulting target text was to be placed in the assignment wiki page. About 1–2 weeks were allocated for each assignment and about 3–4 assignments were given per semester. After the submission and grading of each assignment, the translated texts from each group were opened for all groups to view and were discussed in class.

At the end of each semester we conducted a survey to assess student attitudes toward the use of TransWiki. These focused on two main areas: (1) collaborative learning and (2) the TransWiki system. Below are the ten survey statements used in the most recent (April 2012) survey.

1. Doing assignments in a group helped improve the quality of translation
2. Discussing assignments exposed me to different ways in which language is used across regions/countries
3. I was able to freely express opinions/disagreement during discussions
4. My group managed to arrive at consensus through discussion
5. I would prefer working on assignments individually
6. Using a web-based platform for discussions was convenient
7. Transwiki was easy to learn and access
8. The discussion page facilitated discussions
9. I would like a live-chat function for discussions
10. I would like a discussion page that reflects changes instantly and automatically

Statements 1–5 concerned collaborative learning, statements 6–10 concerned the TransWiki system. A total of 27 students responded whose answers are summarized in Table 1. The results show a strong support for collaborative translation (statements 1 and 2, 81% and 65% support, respectively). Even when asked whether students preferred working individually rather than collaboratively (statement 5), only about one third of students expressed this preference, whereas another third expressed the opposite preference and the remaining students were neutral on this issue. Asked about TransWiki, the majority of students supported it in terms of learning and use (statements 6–8, with 58%, 63% and 50.0% support, respectively). Statements 9–10 asked for feedback regarding the discussion function which had attracted student criticism in the past. As before, students favour a more real-time discussion facility, with the larger

Table 1. TransWiki evaluation results (percentages)

Statement	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
1	11.1	70.4	14.8	3.7	0.0
2	19.2	46.2	26.9	7.7	0.0
3	15.4	65.4	11.5	7.7	0.0
4	12.0	56.0	28.0	4.0	0.0
5	7.7	26.9	30.8	23.1	11.5
6	23.1	34.6	7.7	19.2	15.4
7	25.9	37.0	11.1	25.9	0.0
8	7.7	42.3	26.9	19.2	3.8
9	20.0	40.0	32.0	8.0	0.0
10	50.0	30.8	19.2	0.0	0.0

portion of students expressing a preference for an automatically updating discussion page (81% support). Given this feedback we have in the meantime developed such an instantly updating discussion page using AJAX, similar to the Facebook discussion page, and are deploying it for use in the current semester.

Overall the surveys suggest that working with TransWiki has been generally productive and that the system was well received. Students had complaints regarding the system's speed, especially updating postings on the discussion page. However, use of the discussion page was also appreciated by students who identified lack of peer pressure to agree/conform, as in face-to-face discussions, as one of TransWiki's advantages. The issue of speed was less important where groups allowed discussions to spread out over the entire length of time allowed instead of trying to get together at an appointed time. Such groups generally used the notification function in the wiki which informs students by e-mail once a group member makes a posting. The discussion page is indeed the most important feature of TransWiki as it is at the core of providing a collaborative platform for learning. It also allows all group discussions to be recorded, thus aiding evaluation by the instructor and providing useful insights as to potential problems faced by students. This allows for more focused and individualized teaching.

Regarding disadvantages of the system, a certain portion of students found working with TransWiki cumbersome or difficult. Key areas of complaint included having to log in to the TransWiki page repeatedly, the overall time consumed by the exercise, and discomfort with using computers.

In sum, the use of TransWiki can aid both learning and evaluation. The modus operandi described in this paper also serves as a framework for implementation of constructivist learning environments using TransWiki. The key advantage is the ability to monitor and evaluate the actual collaboration that takes place (if at all). On the other hand ready access to computers and the internet, and being comfortable with using computers over an extended period of time for collaborative assignments are preconditions for successful implementation.

5 Conclusions

We have presented TransWiki, a customized wiki system for use in translation teaching. Our design is based on the existing MediaWiki system, extended with custom code developed by us. In the translation community it is well established that collaborative translation produces better quality results than individual translation. The challenge, however, lies in actually bringing about collaboration among a group of translators. Our experience with our students confirms that TransWiki facilitates online collaboration and discussion on translation. Through TransWiki students feel free to express their opinions and to collaborate, which is particularly relevant for the Asian context where students often hold back their opinions in face-to-face settings but feel more free to express themselves online. TransWiki's features enable instructors to provide separate working areas for separate student groups, enabling focused group work. Moreover, having a digital record not only of the finished translations but also of the communication within each group that led to the translation allows instructors to review and assess both process and outcome of translation, and to identify how learning has occurred. Thus TransWiki is both an enabling platform for collaborative translation and a pedagogical tool for translation teaching.

The design and implementation of TransWiki are not limited to use in translation, however. Other online collaborative work that involves joint construction of content should equally benefit from the features provided by TransWiki.

References

1. Alavi, M.: Computer-mediated collaborative learning: an empirical evaluation. MIS Quarterly 18(2), 159–174 (1994)
2. Leuf, B., Cunningham, W.: The Wiki Way: Quick Collaboration on the Web. Addison-Wesley Longman Publishing Co., Inc. (2001)
3. Pusey, P., Meiselwitz, G.: Heuristics for Implementation of Wiki Technology in Higher Education Learning. In: Ozok, A.A., Zaphiris, P. (eds.) OCSC 2009. LNCS, vol. 5621, pp. 507–514. Springer, Heidelberg (2009)
4. Schwartz, L., Clark, S., Cossarin, M., Rudolph, J.: Educational wikis: features and selection criteria. The International Review of Research in Open and Distance Learning 5(1) (2004)
5. Bruns, A., Humphreys, S.: Wikis in teaching and assessment: the m/cyclopedia project. In: Proceedings of the 2005 International Symposium on Wikis, pp. 25–32. ACM, New York (2005)
6. Lund, A., Smørdal, O.: Is there a space for the teacher in a wiki? In: Proceedings of the 2006 International Symposium on Wikis, pp. 37–46. ACM, New York (2006)
7. Wang, C.M., Turner, D.: Extending the wiki paradigm for use in the classroom. In: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC 2004), vol. 2, IEEE Computer Society, Washington, DC (2004)

8. Hampel, T., Selke, H., Vitt, S.: Deployment of simple user-centered collaborative technologies in educational institutions - experiences and requirements. In: Proc. 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise, pp. 207–214. IEEE Computer Society, Washington, DC (2005)
9. Xu, L.: Project the wiki way: using wiki for computer science course project management. J. Comput. Sci. Coll. 22(6), 109–116 (2007)
10. Icaza, J., Heredia, Y., Borch, O.: Project oriented immersion learning: Method and results. In: 6th International Conference on Information Technology Based Higher Education and Training (ITHET 2005), pp. T4A–7. IEEE (2005)
11. Wang, H.C., Lu, C.H., Yang, J.Y., Hu, H.W., Chiou, G.F., Chiang, Y.T.: An empirical exploration of using Wiki in an English as a second language course. In: Proc. Fifth IEEE International Conference on Advanced Learning Technologies, pp. 155–157. IEEE Computer Society, Washington, DC (2005)
12. Bower, M., Woo, K., Roberts, M., Watters, P.: Wiki pedagogy–a tale of two wikis. In: 7th International Conference on Information Technology Based Higher Education and Training (ITHET 2006), pp. 191–202. IEEE (2006)
13. Biasutti, M., El-Deghaidy, H.: Using wiki in teacher education: Impact on knowledge management processes and student satisfaction. Comput. Educ. 59(3), 861–872 (2012)
14. Šerbec, I.N., Strnad, M., Rugelj, J.: Assessment of wiki-supported collaborative learning in higher education. In: Proceedings of the 9th International Conference on Information Technology Based Higher Education and Training, pp. 79–85. IEEE Press, Piscataway (2010)
15. Tsai, W.T., Li, W., Elston, J., Chen, Y.: Collaborative learning using wiki web sites for computer science undergraduate education: A case study. IEEE Trans. on Educ. 54(1), 114–124 (2011)
16. Tselios, N., Altanopoulou, P., Katsanos, C.: Effectiveness of a framed wiki-based learning activity in the context of HCI education. In: Proceedings of the 2011 15th Panhellenic Conference on Informatics, pp. 368–372. IEEE Computer Society, Washington, DC (2011)
17. Ebner, M., Kickmeier-Rust, M., Holzinger, A.: Utilizing wiki-systems in higher education classes: a chance for universal access? Univers. Access Inf. Soc. 7(4), 199–207 (2008)
18. Raitman, R., Ngo, L., Augar, N.: Security in the online e-learning environment. In: Proc. Fifth IEEE International Conference on Advanced Learning Technologies, pp. 702–706. IEEE Computer Society, Washington, DC (2005)
19. Elrufaie, E., Turner, D.A.: A wiki paradigm for use in IT courses. In: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC 2005), vol. 2, pp. 770–771. IEEE Computer Society, Washington, DC (2005)
20. Tudge, J.: Vygotsky, the zone of proximal development, and peer collaboration: Implications for classroom practice. In: Moll, L.C. (ed.) Vygotsky and Education: Instructional Implications and Applications of Sociohistorical Psychology, pp. 155–172. Cambridge University Press (1992)
21. Király, D.C.: A social constructivist approach to translator education: Empowerment from theory to practice. St. Jerome Pub. (2000)

Physicians' Adoption of Electronic Medical Records: Model Development Using Ability – Motivation - Opportunity Framework

Rajesri Govindaraju, Aulia F. Hadining, and Dissa R. Chandra

Information System and Decision Laboratory, Industrial Engineering Faculty,
Institut Teknologi Bandung Labtek III Building, 4th Floor, Ganesa 10 Bandung 022-2508141
rajesri_g@mail.itb.ac.id, {aulia.fasha,dissarc}@gmail.com

Abstract. The benefits of electronic medical record (EMR) adoption by medical personnel, such as physicians, and medical organizations have been discussed in previous studies. However, most of medical personnel and organizations still use traditional paper-based medical records or use EMR ineffectively. This study aims to develop a model of EMR adoption among physicians and analyse the factors influencing the adoption. The model is developed base on Ability, Motivation, and Opportunity (AMO), adapted AMO, and Motivation-Ability FrameworkTen hypotheses were developed in this study. The next part of the study will be done to operationalize and empirically test the model using a survey method.

Keywords: EMR, Adoption, AMO, physicisians.

1 Introduction

In article 46 paragraph 1 of Medical Practice Law [21], a medical record is defined as "file that contains records and documents about a patient's identity, and also medical examinations, treatments, actions, and other services provided to the patient". Medical records can be used to help physicians in documenting historical records and patient service management [17], [22]. Compared to paper-based medical records, electronic medical records (EMR) give a greater possibility for physicians to improve their work performance quality [17]. The impact of the use of EMR is also mentioed in [13] which stated that 64.3% of studies on EMR found that EMR can improve the performance of medical personnel.

Although benefits of EMR have been scientifically discussed by many studies, most of medical personnels use traditional paper-based medical records or use EMR ineffectively. It may be attributed to the failure in EMR adoption [14] that is related to users' motivation, which consists of internal and external factors [23]. Considering the low adoption rate of EMR among medical personnels, the study reported here aims at developing a model that helps to explain the adoption of EMR by physicians. Although some studies had been done on EMR adoption, most of the respondents in previous studies are medical personnel or medical organizations that have knowledge

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 41–49, 2013.

and have used EMR in some or all processes of their services [8], [13], [14], [17], [22]. A small number of studies discussed EMR adoption in the preparation stage with personnel or medical organizations who have not adopt EMR as their respondents [4], [6], [11], [19]. Further, most of the medical personnel in previous studies are not physicians [3], [8], [9], [13], [24]. Considering the gap in literature mentioned above, this study aims at developing a model of EMR adoption among physicians with physicians who have adopted and have not yet adopted EMR as respondets and analyze the factors influencing the adoption.

2 Model Development

2.1 Ability, Motivation, and Opportunity Theory

Ability, Motivation, and Opportunity (AMO) theory explains that information processing by a person depends on his motivation, opportunity and ability [10]. In this theory, "Motivation" influences "Behavior" with "Ability" and "Opportunity" as moderating influencing factors. The relationships are presented in Figure 1.

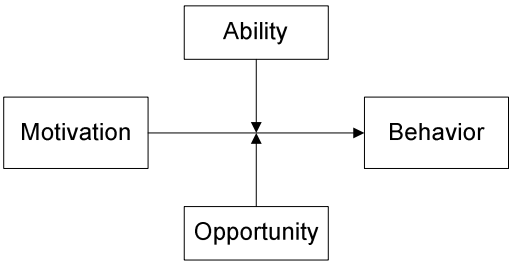


Fig. 1. Ability, Motivation, and Opportunity Theory [10]

Introduced to model brand information processing from advertisements, the variables in the model were defined in the context of advertisement. Ability is defined as skill or proficiency in interpreting brand information in an advertisement [15]. Motivation is defined as the desire to process brand information in the advertisement [15]. Opportunity reflects the extent to which circumstances evidenced during advertisement exposure are favorable for brand processing [15].

Firstly used in studies on information processing such as study on new product information processing [7] and extended study on brand information processing [16], AMO framework has been used in many studies in different areas of research. In the innovation adoption area, AMO was among others used to understand the individual behavior in the online commerce adoption [23]. AMO was also used to discuss the public health and social issue behaviors [20] and consumer behavior [18].

2.2 Adapted Ability, Motivation, and Opportunity Theory

Among the studies that are based on AMO, there are studies that proposed "Intention" as variable that mediates the influence of "Motivation" on "Behavior" (e.g. Olander & Thogersen, 1995). Further, [18] used "Ability" and "Opportunity to moderate the influence of intention on behavior. The adapted AMO framework developed by Olander & Thogersen (1995) is presented in Figure 2.

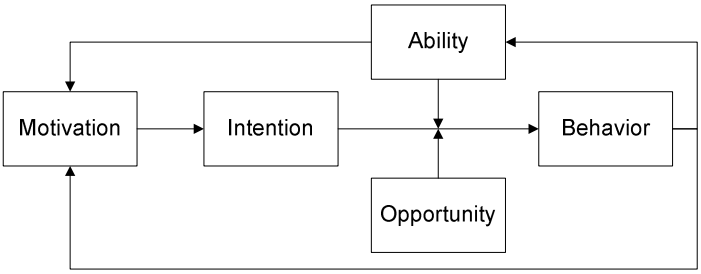


Fig. 2. Adapted Ability, Motivation, and Opportunity Framework [18]

2.3 EHRS Adoption Model

Anderson et.al [2] stated that in an innovation adoption, individual behavior is influenced by ability and motivation. The Electronic Health Record System (EHRS) Adoption Model developed by Anderson et.al [2] is presented in Figure 3. In the case of technology adoption, ability is related to a person's information system skill and is

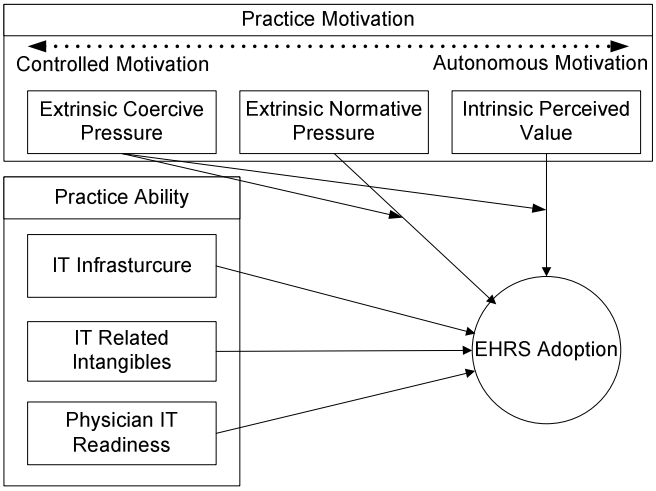


Fig. 3. EHRS Adoption Model [2]

defined as "Practice ability". In the model, "Practice Ability" consists of "IT infrastructure", "IT related intangibles", and "Physician IT readiness". In addition to "Practice ability", "Practice Motivation" is argued to have an influence on EMR adoption. Motivation pushes a person to give responds. The respond possibilities are willingness to adopt or to not adopt EMR. Motivation sources are external and internal factors. Motivation from external factors, which is called extrinsic motivation, is classified into "Extrinsic coercive pressure" and "Extrinsic normative pressure". Meanwhile, intrinsic motivation from internal factors is called "Intrinsic perceived value".

2.4 Research Model Development

Based on AMO framework, Adapted AMO, and EHRS Adoption Model, this study developed a model of EMR adoption among physicians. In general, AMO was used as the main theoretical foundation in building the research model. In this study "Motivation", "Opportunity", and "Ability" were defined as physicians' motivation, opportunity, and ability to adopt EMR. AMO theory is used as a foundation in building the model because it is believed that in the context of EMR adoption, "Motivation" is considered as the drives, urges, wishes or desires which initiate the physician intention to use EMR [15]. "Ability" can facilitate physician in performing adoption of EMR and "Opportunity" is interpreted as situational factors that encourage physician adoption of EMR.

"Behavior" in MOA was changed into "EMR adoption". "Behavior" can be achieved through "Intention" which is controlled by conscious motivation by the physician [1]. Thus, "Intention" was added to mediate the relationship between "Motivation" and "Behavior"/"EMR Adoption". It is in line with the concept used in [10] and adapted MOA model used in [18], which are originated from the Theory of Planned Behavior (TPB). TPB stated that "Intention" is indication of how hard people are willing to try in order to perform a behavior [1]. In line with [18], we argue that "Ability" and "Opportunity" moderate the influence of "Intention" on "Adoption". "Ability" is defined as the capabilities that facilitate physician in performing adoption of EMR whereas "Opportunity" is defined as situational factors that encourage physicians to adopt EMR.

As presented in Figure 4, there are 10 variables and 3 variable groups in the model. In the next part, the development of the hypoteses will be discussed. The discussion will be presented in the following four different parts:

1) Intention.

The first hypothesis in this study is that "Intention" to use EMR will have an impact on "EMR Adoption" (behavior).

H1: Physicians' "Intention" to use EMR positively influences their "EMR adoption".

2) Motivation.

The first variable group is "Motivation" which is adapted from [15] and [2]. In this model, "Motivation" consists of "Extrinsic coercive pressure", "Extrinsic normative

pressure", and "Intrinsic perceived value". "Extrinsic normative pressure" is is normative in nature and addresses the question of how many other practices with which the focal practice routinely interacts have already adopted EMR [2]. In a study of interorganizational linkages adoption, normative pressure exhibited the strongest influence on organization-level technology adoption" [2]. Intrinsic motivation from internal factors is called "Intrinsic perceived value". Intrinsic motivation involves an individual acting out of an internal belief that the activity is interesting, good, satisfying or right. Intrinsic motivation is inherently autonomous and the behavior showed by it is characterized by individual choice and volition [2]. This form of motivation tends to yield positive outcomes in terms of job satisfaction, effective performance and feelings of competence [2].

H2: "Extrinsic normative pressure" felt by physicians positively influences their "Intention".

H3: Physicians' "Intrinsic motivation" positively influences their "Intention".

A physician will be eager to adopt EMR, if he often interacts with external stakeholders who also use EMR [2], [12], [23], [25]. As an example is a situation in a clinic in which medical services have already been done with the support of EMR. Physicians who work at the clinic receive pressure to adapt and follow the technology development. In this situation, the physicians are usually more cooperative to adopt EMR.

H4: "Extrinsic normative pressure" positively moderates the influence of "Intrinsic perceived value" on "Intention" to use EMR.

"Extrinsic coercive pressure" is defined as pressure to adopt EHRS from external entities [2]. The forceful characteristic of "Extrinsic coercive pressure", which examples are government laws, hospital regulations, and medical association standards, influences physicians to respond negatively to the pressure itself [2], [5], [17]. In line with [2], this research also addresses interactive relationship between motivating factors. "Extrinsic coercive pressure" may be a motivating factor, but this motivation is a form of controlled motivation that undermines the positive influence of intrinsic motivation. It means that "Extrinsic coercive pressure" is a factor that reduces the influence of physicians' "Intrinsic perceived value" on "intention" to use EMR [2].

H5: "Extrinsic coercive pressure" negatively moderates the influence of "Intrinsic perceived value" on "Intention" to use EMR.

The influence of "Extrinsic normative pressure" on "EMR adoption" is also influenced negatively by "Extrinsic coercive pressure" [2]. It is based on the assumption that the "Extrinsic coercive pressure" reduces the positive influence of "Extrinsic normative pressure" in a similar way it influences the positive influence of "Intrinsic perceived value" on "EMR adoption".

H6: "Extrinsic coercive pressure" negatively moderates the influence of "Extrinsic normative pressure" on "Intention" to use EMR.

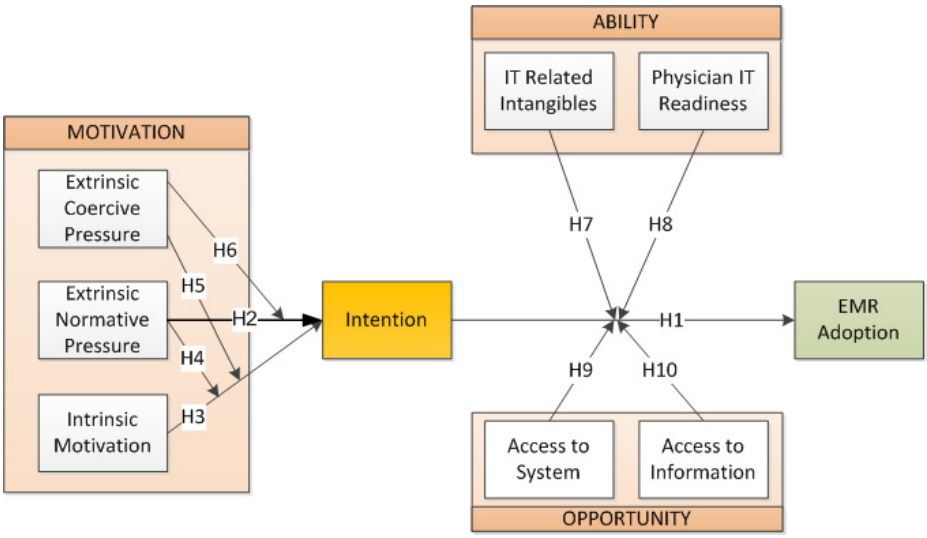


Fig. 4. EMR Adoption Model

3) Ability.

"Intention" to adopt EMR has to be supported by sufficient "Ability" and "Opportunity" in order to establish a behavior of EMR use. In this research model, "Ability" consists of "IT related intangibles" and "IT readiness". "IT related intangibles" is defined as a condition that users have already used IT and obtained benefits [2]. The last component is "Physician IT readiness". This variable is defined as a condition in which existing knowledge gives physicians responsive and agile abilities in adopting a new technology [2]. As explained in MOA, "Ability" moderates the relationship between "Intention" and "Behavior". In line with that, in this research model, the two components of "Ability" also influence the relationship between "Intention" and "Behavior". The related hypotheses are stated below.

H7: "IT related intangibles" positively moderates the influence of "Intention" to use EMR" on "EMR adoption".

H8: "Physician IT readiness" positively moderates the influence of "Intention" to use EMR on "EMR adoption".

4) Opportunity.

"Opportunity" is an uncontrollable aspect for a person [7]. "Opportunity" in this study is also reffered to as aspects that could encourage adoption of EMR, indirectly. "Opportunity" can be defined as external conditions and situations that cannot be controlled by a person and moderate the desires to adopt EMR to occur [7], [15]. The general ideas of "Opportunity" in previous studies are: (1) the contact between the subject, which is the customer, and the object, which is the advertisement; and (2) the object's characteristics. In the context of public relation message processing behavior studied in [7], examples for the first type opportunity are exposure time and the absence of distractions that detract from message processing, and examples for the

second type opportunity are message length and the number of arguments. In the context of EMR adoption, physicians' opportunity is not only about their contact with EMR system. Physicians always have opportunity to use EMR as long as the organization's EMR system is accessible. If the system is accessible, they may have contacts with the system directly or indirectly through the peer practices. Further, the accessibility of EMR system is also affected by its' characteristics, such as the IT infrastructure and the system functions. Therefore, the first variable in "Opportunity" is "Access to system". In addition, in this model "Access to information" is included as an opportunity variable. This variable represents physicians' opportunity to get in contact with any information media or other sources. However, the opportunity is not affected by the physicians' available time. "Access to information" means access to information from which the physicians can improve their ability and knowledge, as well as their motivation. Examples of the information sources are training, knowledge sharing forum, and information media. Thus, in this model, "Opportunity" consists of "Access to system" and "Access to information". Both the factors support the establishment of "EMR Adoption" from the existing "Intention".

H9: "Access to System" positively moderates the influence of "Intention" to use EMR on "EMR adoption".

H10: "Access to Information" positively moderates the influence of "Intention" to use EMR on "EMR adoption".

3 Methods and Discussion

This paper presents a preliminary study to develop a model of EMR adoption. In the next step, the defined hyphoteses will be tested using an empirical. Before collecting data, the model will be operationalized based on earlier published literature, and the measured variables will be defined. Then, data collection process will be done using a questionnaire survey method. The quantitative data will be analyzed using statistical approach with Partial Least Square and resulting qualitative data will be used to infer the results of research. The sampling method in this study will be purposive sampling in which sampe will be chosen considering the study's objective.

This study's objective is to develop a model of EMR adoption among EMR physicians and analyse the factors influencing the adoption. However, as discussed in the model development, there are two extrinsic motivation variables in the model. As we want to get a better understanding about these variables, this study will be focused on EMR adoption by physicians who join in medical organizations. In this study the empirical data will be collected from physicians working in small group practices of twenty or fewer physicians, in which EMR use are encouraged, but not mandatory. We use small group practices because this form of medical service arrangement is popular in Indonesia. Questionnaires will be sent to physicians who join these small group practices providing. These small group practices usually facilitate the physicians with EMR systems to support the administration of the health care services provided to the patients.

References

1. Ajzen, I.: The Theory of Planned Behavior. Organizational Behavior and Human Decision Processes, 179–211 (1991)
2. Anderson, C.L., Mishra, A.N., Agarwal, R., Angst, C.: Digitizing Healthcare: The Ability and Motivation of Physician Practices and Their Adoption of Electronic Health Record Systems. In: Twenty Eighth International Conference on Information Systems, pp. 1–17 (2007)
3. Boonstra, A., Broekhuis, M.: Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Medical Informatics and Decision Making 10, 1–17 (2010)
4. Cauldwell, M., Beattie, C., Cox, B., Denby, W., Ede-Golightly, J., Linton, F.: The Impact of Electronic Patient Records on Workflow in General Practice. Health Informatics Journal 13, 155–162 (2007)
5. Ford, E.W., Alearney, A.S.M., Phillips, M.T., Menachemi, N., Rudolphe, B.: Predicting Computerized Physician Order Entry System Adoption in US Hospitals: Can The Federal Mandate be Met? International Journal of Medical Informatics 77, 539–546 (2008)
6. Garets, D., Davis, M.: Electronic Medical Records VS Electronic Health Records: Yes, There Is a Difference. HIMSS Analytics, 1–14 (2006)
7. Hallahan, K.: Enhancing Motivation, Ability, and Opportunity to Process Public Relations Messages. Public Relations Review 26, 463–481 (2000)
8. Hennington, A., Janz, B.D.: Information Systems and Healthcare XVI: Physician Adoption of Electronic Medical Records: Applying the UTAUT Model in a Healthcare Context. Communications of the Association for Information Systems 19, 60–82 (2007)
9. Heselmans, A., Aertgeerts, B., Donceel, P., Geens, S., Velde, S.V., Ramaekers, D.: Family Physicians' Perceptions and Use of Electronic Clinical Decision Support During the First Year of Implementation. J. Med. Syst. 8 (2012)
10. Hughes, J.: The Ability – Motivation - Opportunity Framework for Behavior Research in IS. In: Proceedings of the 40th Hawaii International Conference on System Sciences - 2007, vol. 7, p. 10 (2007)
11. Hunter, I.M., Whiddett, R.J., Norris, A.C., McDonald, B.W., Waldon, J.A.: New Zealanders' Attitudes Toward Access to Their Electronic Health Records: Preliminary Results From a National Study Using Vignettes. Health Informatics Journal, 212–227 (2009)
12. Ilie, V., Courtney, J.F., Slyke, C.V.: Paper versus Electronic: Challenges Associated with Physicians' Usage of Electronic Medical Records. In: Proceedings of the 40th Hawaii International Conference on System Sciences, vol. 10 (2007)
13. Lau, F., Price, M., Boyd, J., Partridge, C., Bell, H., Raworth, R.: Impact of electronic medical record on physician practice in office settings: a systematic review. BMC Medical Informatics and Decision Making 12, 10 (2012)
14. Ludwick, D.A., Doucette, J.: Adopting Electronic Medical Records in Primary Care: Lessons Learned From Health Information Systems Implementation Experience In Seven Countries. International Journal of Medical Informatics 78, 22–32 (2009)
15. MacInnis, D.J., Jaworski, B.J.: Information Processing form Advertisements: Toward an Integrative Framework. Journal of Marketing 53, 1–23 (1989)
16. Maclnnis, D.J., Moorman, C., Jaworski, B.J.: Enhancing and Measuring Consumers' Motivation, Opportunity, and Ability to Process Brand Information From Ads. Journal of Marketing 55, 32–55 (1991)
17. Miller, R.H., Sim, I.: Physicians' Use Of Electronic Medical Records: Barriers And Solutions. Health Affairs, 116–127 (2004)

18. Olander, F., Thogersen, J.: Understanding of Consumer Behaviour as a Prerequisite for Environmental Protection. Journal of Consumer Policy 18, 345–386 (1995)
19. Randeree, E.: Exploring Physician Adoption of EMRs: A Multi-Case Analysis. [Original Paper] Springerlink 8 (2007)
20. Rothschild, M.L.: Carrots, Sticks, and Promises: A Conceptual Framework for the Management of Public Health and Social Issue Behaviors. Journal of Marketing 63, 24–34 (1999)
21. Sjamsuhidajat, Alwy, S., Rusli, A., Rasad, A., Enizar, Irdjiati, I., et al.: Manual Rekam Medis. Jakarta Selatan, Konsil Kedokteran Indonesia (2006)
22. Su, Y.Y., Win, K.T., Chiu, H.C.: Development of Taiwanese Electronic Medical Record Systems Evaluation Instrument. International Journal of Biological dan Life Sciences, 140–145 (2008)
23. Teh, P.-L., Ahmed, P.K.: MOA and TRA in Social Commerce: An Integrated Model. In: Proceedings of the 2011 IEEE IEEM , vol. 11, pp. 1375–1379 (2011)
24. Walter, Z., Lopez, M.S.: Physician acceptance of information technologies: Role of perceived threat to professional autonomy. Decision Support Systems 46, 206–216 (2008)
25. Wills, M.J., El-Gayar, O.F., Bennett, D.: Examining Healthcare Professionals' Acceptance of Electronic Medical Records Using UTAUT. Issues in Information Systems IX, 396–402 (2008)

Software Development Methods in the Internet of Things

Selo Sulistyo

e-Systems Lab
Department of Electrical Engineering and Information Technology
Gadjah Mada University
Jl. Grafika No. 2, Bulaksumur, Yogyakarta, Indonesia 55281
selo@ugm.ac.id

Abstract. In the Internet of Things, billions of networked and software-driven devices will be connected to the Internet. They can communicate and cooperate with each other to function as a composite system. This paper proposes the AMG (abstract, model and generate) method for the development of such composite systems. With AMG, the development of software application can be done in an automatic manner, and therefore reducing the cost and develpment time. The method has been prototyped and tested with use cases.

1 Introduction

Today's Internet technology is mainly built for information sharing. Information providers, which typically are implemented as servers, provide information in the form of web pages that can be accessed by internet clients. In the *Future Internet* [16], various independent networked computing devices from small devices (mobile devices, embedded systems, etc) to powerful devices (desktops and servers) may be easily connected to the Internet, in a plug and play manner. The Internet of Things is one of the popular terms illustrating the *Future Internet*.

From the software developer's point of view, the '*Thing*' in the Internet of Things can be seen as all kinds of networked devices that are driven and delivered by (embedded) software. Considering that the device's functionalities are provided as services, we will have billions of services in the Internet (i.e., the Internet of Services). A typical example of an environment containing several (embedded) services is a smart home where a residential gateway is controlling and managing home devices with embedded services. In this type of dwelling, it is possible to maintain control of doors and window shutters, valves, security and surveillance systems, etc. It also includes the control of multi-media devices that are parts of home entertainment systems. In this scenario, the smart home is containing 1*) WeatherModules that provide different data collection services (i.e., air temperature, solar radiation, wind speed, and humidity sensors)*, 2*) Lamps that provide on-off and dimmer services*, and 3*) Media Renderers that provide playing of multimedia services*, see Figure 1 below.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 50–59, 2013.

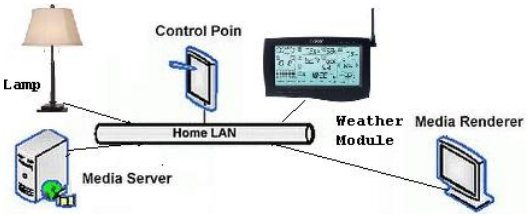


Fig. 1. Smart home services: A Scenario

Combining these independent services (i.e., embedded software) and promoting as a new application is a challenge. Unfortunately, traditional software engineering approaches are not fully appropriate for the development of service-based applications. There is an urgent need for developing comprehensive engineering principles, methodologies and tools support for the entire software development lifecycle of service-based applications [19].

This paper proposes a software development method in the Internet of Things. The remainder of the paper is organized as follows: In Section 2 we present terms and definitions of a service and a service-based application. Then, in Section 3 we present the AMG method. Section 4 is devoted to related work. Finally, we draw our conclusions in Section 5.

2 Background

This section gives a background for the paper. It discusses the definition of a service and a service-based application.

What Is a Service? A service can be defined in different ways. In [17] for instance, a service is defined as *asset of functions provided by a (server) software or system to client software or system, usually accessible through an application programming interface.* Similar definitions as the one above appear in the context of middleware technologies such as Jini [3], .NET [9], or JXTA [21]. These definitions recognize services as a central element in the system implementation.

In this paper, referring to [1] and [10] a service is defined as a model of software unit. To get an overview of this definition, we have to look at the history of managing the complexity of software systems where a component-oriented architecture is considered as the solution for the complexity problem. Figure 2 illustrates a historical perspective of the use of different models to represent a software unit.

The idea of using software component as defined in [10], can be considered as the birth of today's software component and can be seen as an architectural approach of building software systems. When Assembler was the only available programming language, a routine was considered the first model of software units. As the complexity of software systems was increasing a new model of a

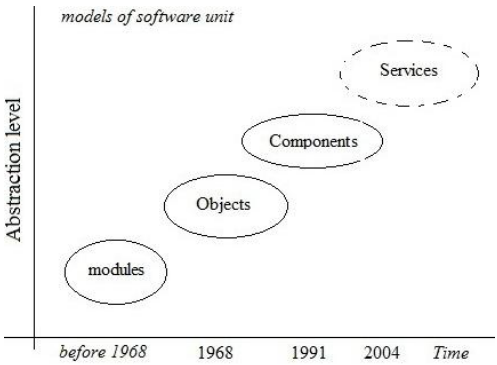


Fig. 2. The historical perspective of the use of different models of a software unit

software unit called a module was introduced. A module is a simple model. However, a module is more abstract than a routine. Accordingly, objects, components and services can also be seen as models of a software unit. The differences between them are their abstraction levels, means of encapsulation and ownership. Software abstraction, encapsulation and reuse are the key points.

With software component-orientation in mind, a single software component might not work as an application. Therefore a composition system is needed. According to [1], a software composition system has three aspects: component models, composition techniques and composition languages. Depending on the models of software unit, see Figure 2, different composition techniques and languages are required. These two aspects have influenced the development approaches and paradigms. For example, when we use objects as a model of a software unit to build a software system we call the paradigm object-oriented development. Accordingly, we have component-oriented development for component oriented-systems and service-oriented development for service-oriented systems. A general concept of service-oriented systems is well-known as Service-Oriented Architectures (SOA)[4].

Service-Based Applications. Using the SOA concept, architecturally, software applications are built from compound, heterogeneous, autonomous software units called services. If it is the case, service compositions will be a common approach for the development of software applications in the Internet of Services. Software systems and applications are becoming service-based. We call this a service-based application.

A conceptual model of a service-based application is presented in Figure 3. It can be seen that a service-based application may use more than one service. It is shown also that a service can be classified either as `Simple` or `Composite`. In this paper, all services that will be composed are considered as a simple service. As mentioned earlier, a composite service is a type of a service-based application. It composes services and provides the combined functionality as new services.

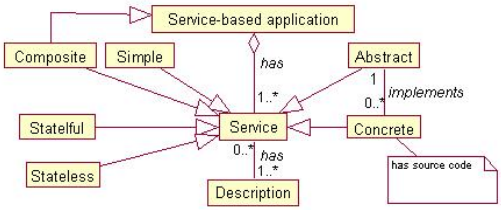


Fig. 3. A conceptual model of service-based applications

A service can also be classified either as `Abstract` or `Concrete`. An abstract service may have one or more concrete services or it may mean that the service will be implemented in the future. But for the composition of services, this paper considers only Concrete services (run-time services). Furthermore, services can also be classified either as a `State-less` or `State-ful` service. Web services can be considered as an example of stateless services, while UPnP services [8] can be considered as a kind semi state-ful services. UPnP devices use state variables to store the states of specific variables and inform those state changes to other UPnP devices.

3 The AMG Method

With regard to the software production, there are different approaches, which focus on how to specify, design, implement, test, and deploy software systems. They can be categorized as implementation- and model-oriented approaches. AMG is a model-oriented approach. Figure 4 shows the idea of the AMG-method.

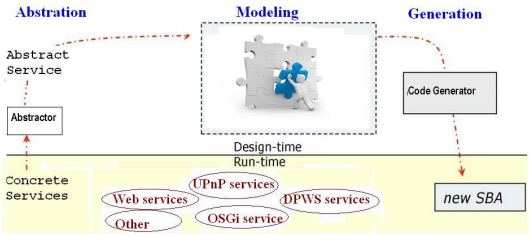


Fig. 4. The AMG (abstract, model, and generate) method

3.1 The Abstraction Step

Specifying models of a service-based application is only possible if the models of the included (i.e., existing) services are in place. For this an abstraction process is required. The abstraction step consists of a transformation mechanism of service descriptions into graphical representations and source code.

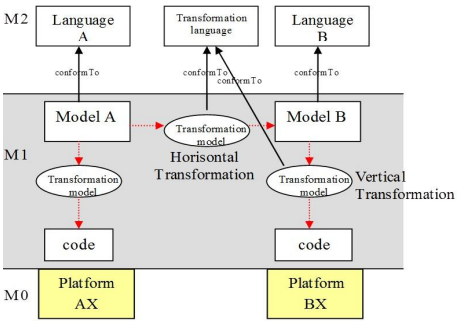


Fig. 5. A conceptual model of model transformation

The abstration step uses the concept of model transformations [13]. Figure 5 illustrates the concept of model transformation. As shown in the figure, two main types of transformation exists; vertical transformation and horizontal transformation. In the vertical transformation, a source model is at a different level of abstraction than a target model. Examples of vertical transformation are refinement (specialization), PIM-to-PSM transformations and abstraction (generalization). Generalization could mean also an abstraction of platforms or a transformation from code into models.

In tge horizontal transformation, the source model has the same level of abstraction as the target model. Examples of the horizontal transformation are refactoring and merging. In this type of transformation one or more source models are transformed into one or more target models, based on the languages (meta-model) of each of these models. In other words the instance of one meta-model is transformed into instances of another meta-model. So, in this step, we apply the vertical transformation.—

Models can be presented in two forms; graphical (models) or textual. Depending on the relation between these form, there are four different model transformations; M2M (Model-to-Model) transformation, T2M (Text-to-Model) transformation, M2T (Model-to-Text) transformation and T2T (Text-to-Text) transformation. The T2T transformation is often used for the processing of the M2M, M2T and T2M.

Using the illustrated concept earlier, from a service description (s), the abstractor produces a graphical service model ($\mathbf{M}_s$) conforming to a selected modeling language and source code ($\mathbf{C}_s$) conforming to a selected programming language. To automate the transformation process, existing service frameworks and APIs (e.g., the Web Service framework and API) are used.

Depending on the selected modeling language, different graphical representations (i.e., notations) can be used to represent the existing services. UML classes, CORBA components, Participants in SoaML, or SCA components are among them. However, it must be noted that within the context of domain-specific modeling (DSM), a graphical representation must relate to a real thing which in this case is the implementation of the service. Therefore, it is important to keep

the relation (bindings) between graphical representations (i.e. service models) and source code (i.e. implementation for the service invocations). Fig. 6 shows the relation between a service description, its model, and source code.

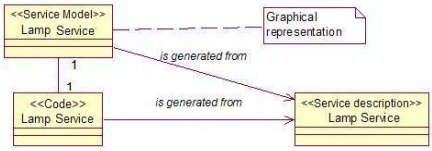


Fig. 6. The relation between a service description, its model and source code

Different service frameworks and APIs have been developed using different programming languages and run on different platforms, helping developers to implement services. For example, in the Web services context there are Apache Axis (Java and C++), appRain (PHP), .NET Framework (C#, VB and .NET), etc. Therefore, a graphical representation of a service may have several implementations (i.e., source code). This source code may also use different programming languages and may run on different platforms.

An Example: *UML Classes.* A UPnP device has two kinds of descriptions; device description and service description. A UPnP device can have several services that are in a UPnP service description called Actions. To automate the abstraction step we use transformation rules. Table below shows transformation rules to transform different properties in a UPnP service description into properties in an UML class. To construct the transformation rules, both UML and UPnP meta-models are required. However, the rules are very simple. For example, to present the class name, we use the name of the UPnP device. Obviously, other XML-based service descriptions (e.g., WSDL, DPWS) will use the similar process.

3.2 The Modeling Step

In software development, models are used to describe the structures and the behaviors of the software systems. The structure specifies what the instances of the model are; it identifies the meaningful components of the model construct and relates them to each other. The behavioral model emphasizes the dynamic behavior of a system, which can essentially be expressed in three different types of behavioral models; interaction diagrams, activity diagrams, and state machine diagrams. We use interaction diagrams to specify in which manner the instances of model elements interact with each other(roles)

Even though for model-driven development, state-machine diagrams are considered as the most executable models, we are still interested in using UML activity diagram and collaboration diagram. The reason is that from activity

diagrams we can generate state machine diagrams [11]. The UML activity diagrams are used mostly to model data/object flow systems that are a common pattern in models. The activity diagram is also good to model the behavior of systems which do not heavily depend on external events.

AMG is a language-independent method. It is possible to use different existing modeling languages and different modeling editors. This is done by developing and implementing different service abstractors/presenters. The requirement is that the presenter must generate notations (i.e., abstract service models) that conform to the chosen modeling languages. Using the abstract service models $(\mathbf{M}_{s1}, \mathbf{M}_{s2}, ..., \mathbf{M}_{sk})$, a service-based application can be expressed in a composition function f$\{\mathbf{M}_{s1}, \mathbf{M}_{s2}, ..., \mathbf{M}_{sk}\}$, where $s1..sk$ are the included services in the service-based application.

An Example: *UML Sequence Diagram.* Using UML classes, the structure of a service-based application can be specified as a class diagram, while the behavior part can be defined using sequence diagram. Accordingly, the semantic follows the semantic of UML 2.0. Fig. 7 shows a UML model of the service-based application defined in the scenario. There are four UML classes that represent different existing services mentioned in the scenario.

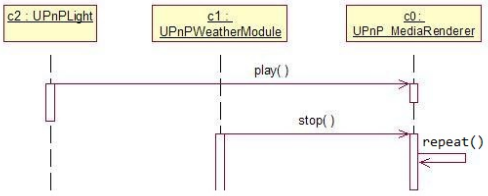


Fig. 7. A service-based application model specified is using a sequence diagram. The composed application will play music when a UPnPLight service is invoked, and will stop it on a certain value of weather parameter.

3.3 The Generation Step

For model execution, we use code generation approaches instead of model interpretations. For this, we did not use any transformation language to generate code, but a Java program to transform models into texts (i.e., source code). The potential code generation from activity diagrams was studied in [2] and [5]. For a tool, Enterprise Architect from Sparx Systems [18] is an example for modeling tools that support code generation from activity diagrams.

The code generation process of a service-based application can be expressed as a generation function g$[f\{\mathbf{M}_{s1}, \mathbf{M}_{s2}, .., \mathbf{M}_{sk}\}, \mathbf{C}_{s1..sk}, dev_info] \Rightarrow code$, where $f\{\mathbf{M}_{s1}, \mathbf{M}_{s2}, .., \mathbf{M}_{sk}\}$ is the model of the service-based application, $s1, s2$ .. sk are the included services, $\mathbf{C}_{s1..sk}$ are the connected code of the used service models $(\mathbf{M}_{s1..sk})$, and dev_info is the given device information (i.e., the capability and configuration information).

Code from the behavior parts is taken from the activity nodes. For this we adapt the generation method presented in [2]. With regard to their method, an UML class can be considered as an entity that executes an external action. For example, for the decision node (i.e., the decision node with the `airtemperature` input) the following code is produces. To generate code from the structure, from each classes, one object is instantiated. Since the UML classes in this scenario are platform independent, the objects to be instantiated are depending on the platform selection.

4 Discussion

Service composition is gaining importance, as it can produce composite services with features not present in the individual basic services. However, the fact that different perspectives may have different definitions of a service, the definition of service composition may also be different. AMG considers a service is just a kind of software component model that has evolved from the older software component models (i.e., modules, objects, and components). With this definition, a services composition can be done in a similar way as a composition of software units that normally is done at design-time using bottom-up approaches.

AMG focuses on service composition at design-time. However, the abstraction step in the AMG can be extended to support run-time compositions. Conceptually, it would be possible to generate graphical service representation that can be used by end-users (i.e., run-time composition). In the ISIS project for example [20], ICE, an end-user composition, has been developed. A service in ICE is presented as a puzzle with either one input (trigger) or one output (action). A composition is done by connecting puzzles. Using a specific service abstractor, ICE puzzles for end-users and its source code for implementing service invocations can be generated.

For the composition of service component models (i.e., software units), composition techniques, and composition languages are required [1]. In Web service context, the Web Service Business Process Execution Language (WS-BPEL) [15] and the Web Services Choreography Description Language (WS-CDL) [6] can be considered as a composition language. Within the OMG context, the Service-oriented architecture Modeling Language (SoaML) [14] is another example of composition languages. Also in the Web services context, services orchestration and choreography are well-known service composition techniques. In the context of Service Component Architecture (SCA) [7], wiring can also be considered as a type of composition techniques. For this reason, AMG method does not depend to any particular languages. This can be done by developing different abstractor for different target languages.

Abstracting software functionality into abstract graphical representation has also been studied by other researchers. For example, in [12] UML classes is used to abstract Web services. However, their work focused only on Web services and did not think other possible services. In [22], software components are visualized using graphical notations that developers can easily understand. They use a

picture of a real device to present a software component. The integration is done by simply connecting components graphically. Obviously, the approach is only applicable for a specific domain. In contrast, the AMG method is domain-independent.

5 Conclusion

With regard to the complexity of software systems, the aims for both software composition and model-driven development (e.g. MDA) are similar in which they are used for managing the complexity of software systems and their development. Having benefited from these approaches, this paper propose the use of model-driven development for the development of software applications in the Internet of Things. With this idea, the composition of services can be done using models at different abstraction levels, while the executable composite services can be generated automatically.

References

1. Assmann, U.: Invasive software composition. Springer (April 2003)
2. Bhattacharjee, A.K., Shyamasundar, R.K.: Validated Code Generation for Activity Diagrams. In: Chakraborty, G. (ed.) ICDCIT 2005. LNCS, vol. 3816, pp. 508–521. Springer, Heidelberg (2005)
3. Cicirelli, F., Furfaro, A., Nigro, L.: Integration and interoperability between jini services and Web services. In: IEEE International Conference on Services Computing, pp. 278–285. IEEE Computer Society, Los Alamitos (2007)
4. Erl, T.: Service-Oriented Architecture (SOA): Concepts, Technology, and Design. Prentice Hall PTR (August 2005)
5. Eshuis, R., Wieringa, R.: A formal semantics for uml activity diagrams - formalising workflow models (2001)
6. Diaz, G., Pardo, J.-J., Cambronero, M.-E., Valero, V., Cuartero, F.: Automatic Translation of WS-CDL Choreographies to Timed Automata. In: Bravetti, M., Kloul, L., Zavattaro, G. (eds.) EPEW/WS-FM 2005. LNCS, vol. 3670, pp. 230–242. Springer, Heidelberg (2005)
7. IBM. Service component architecture, `http://www.ibm.com/developerworks/library/specification/ws-sca/` (November 2006)
8. Jeronimo, M., Weast, J.: UPnP Design by Example: A Software Developer's Guide to Universal Plug and Play. Intel Press (May 2003)
9. JiZhe, L., YongJun, Y.: Research and implementation of lightweight esb with microsoft.net. Japan-China Joint Workshop on Frontier of Computer Science and Technology, 455–459 (2009)
10. McIlroy, D.: Mass-Produced software components. In: Proceedings of the 1st International Conference on Software Engineering, pp. 88–98 (1968)
11. F.A. Kraemer. Engineering Reactive Systems: A Compositional and Model-Driven Method Based on Collaborative Building Blocks. PhD thesis, Norwegian University of Science and Technology, Trondheim (August 2008)

12. Grønmo, R., Skogan, D., Solheim, I., Oldevik, J.: Model-Driven web services development. In: Proceedings of International Conference on e-Technology, e-Commerce, and e-Services, pp. 42–45. IEEE Computer Society, Los Alamitos (2004)
13. OMG. Meta object facility (MOF) 2.0 Query/View/Transformation specification final adopted specification ptc/05-11-01 (2005), http://www.omg.org/docs/ptc/05-11-01.pdf
14. OMG. Service oriented architecture modeling language (SoaML) : Specification for the UML profile and metamodel for services (UPMS) (2009)
15. Ouyang, C., Verbeek, E., van der Aalst, W.M.P., Breutel, S., Dumas, M., Hofstede, A.H.M.t.: Formal semantics and analysis of control flow in WS-BPEL. Science of Computer Programming 67(2-3), 162–198 (2007)
16. Papadimitriou, D.: Future internet: The cross-etp vision document. Technical Report Version 1.0, European Future Internet Assembly, FIA (2009)
17. Sancho. Definition for the term (software) service, sector abbreviations and definitions for a telecommunications thesaurus oriented database, itu-t (2009)
18. Sparx Systems. Enterprise architect, http://www.sparxsystems.com/products/ea/index.html
19. van den Heuvel, W.-J., Zimmermann, O., et al.: Software service engineering: Tenets and challenges. In: Proceedings of the 2009 ICSE Workshop on Principles of Engineering Service Oriented Systems, PESOS 2009, pp. 26–33. IEEE Computer Society, Washington, DC (2009)
20. Su, X., Svendsen, R.M., et al.: Description of the ISIS Ecosystem Towards an Integrated Solution to Internet of Things. Telenor Group Corporate Development (2010)
21. Yeager, W., Williams, J.: Secure peer-to-peer networking: The jxta example. IT Professional 4, 53–57 (2002)
22. Yermashov, K.: Software Composition with Templates. PhD Thesis, De Montfort University, UK (2008)

SAT–Based Bounded Strong Satisfiability Checking of Reactive System Specifications

Masaya Shimakawa, Shigeki Hagihara, and Naoki Yonezaki

Department of Computer Science,
Graduate School of Information Science and Engineering,
Tokyo Institute of Technology,
2-12-1-W8-67 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Abstract. Many fatal accidents involving safety-critical reactive systems have occurred in unexpected situations that were not considered during the design and test phases of the systems. To prevent these accidents, reactive systems should be designed to respond appropriately to any request from an environment at any time. Verifying this property during the specification phase reduces development reworking. This property of a specification is commonly known as realizability. Realizability checking for reactive system specifications involves complex and intricate analysis. For the purpose of detecting simple and typical defects in specifications, we introduce the notion of bounded strong satisfiability (a necessary condition for realizability), and present a method for checking this property. Bounded strong satisfiability is the property that for all input patterns represented by loop structures of a given size k, there is a response that satisfies a given specification. We present a checking method based on a satisfiability solver, and report experimental results.

1 Introduction

A reactive system is one that responds to requests from an environment in a timely fashion. The systems used to control elevators or vending machines are typical examples of reactive systems. Many safety-critical systems, such as those that control nuclear power plants and air traffic control systems, are also considered reactive systems. In designing a system of this kind, the requirements are analyzed and then described as specifications for the system. If a specification has a flaw, such as inappropriate case-splitting, a developed system may encounter unintended situations. Indeed, many fatal accidents involving safety-critical reactive systems have occurred in unexpected situations, which were not considered during the design and test phases of the systems. Therefore, it is important to ensure that a specification does not possess this kind of flaw[8].

More precisely, a reactive system specification must have a model that can respond in a timely fashion to any request at any time. This property is called realizability, and was introduced in [1,12]. In [12], it was demonstrated that a reactive system can be synthesized from a realizable specification. However, realizability checking for reactive system specifications involves complex and

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 60–70, 2013.

intricate analysis. Therefore, the sizes of specifications that can be checked in a practical application are strongly limited.

For the purpose of detecting simple and typical defects in specifications, we present a bounded checking method. In bounded checking, we verify the existence of a counterexample (or witness) of a given size k. Such methods are used successfully in other fields, for example, bounded model checking[3] and Alloy Analyzer[8]. The advantage of bounded checking is its ability to detect a small counterexample (or witness) efficiently.

We present a bounded property for strong satisfiability [10] (a necessary condition for realizability) known as bounded strong satisfiability, together with a method for checking this property. Strong satisfiability is the property that for any input sequences, there is a response that satisfies a given specification. Strong satisfiability can be checked with lower complexity than realizability[16]. Checking strong satisfiability is EXPSPACE-complete. On the other hand, checking realizability is 2EXPTIME-complete. Although this property is a necessary condition, many practical unrealizable specifications are also strongly unsatisfiable[11]. Bounded strong satisfiability restricts the input sequences of strong satisfiability to those represented by loop structures of size k. This means that for simple input patterns, there is a response that satisfies the specification. Our experience has shown that in many instances, strongly unsatisfiable specifications have small counterexamples (which are input sequences that can be represented by loop structures of small size). Thus, we anticipate that many defects can be detected by checking this property.

In our method for checking bounded strong satisfiability, we use an SAT solver. Specifically, we first construct a non-deterministic Büchi automaton (NBA) that accepts input sequences for which there is a response that satisfies a specification. Then we check whether or not the NBA accepts all loop structures of size k, using an SAT solver. To accomplish this, checking the existence of a loop structure that is not accepted by NBA is reduced to an SAT problem. This reduction is based on the following characterization of non-accepted loop structures: A loop structure σ of bounded size is not accepted by NBA if and only if, for any run on σ, final states occur at most d times for some d. This characterization is valid because only bounded loop structures are considered.

We implemented our method, and found that it can handle larger specifications than techniques based on other properties, and can detect defects efficiently. We also report experimental results in this paper.

2 Preliminaries

2.1 Reactive Systems

A reactive system is a system that responds to requests from an environment in a timely fashion.

Definition 1 (Reactive system). *A reactive system RS is a triple $\langle X, Y, r\rangle$, where X is a set of events caused by an environment, Y is a set of events caused by the system, and $r : (2^X)^+ \mapsto 2^Y$ is a reaction function.*

We refer to events caused by the environment as 'input events,' and those caused by the system as 'output events.' The set $(2^X)^+$ is the set of all finite sequences of sets of input events. A reaction function r relates sequences of sets of previously occurring input events with a set of current output events.

2.2 A Language for Describing Reactive System Specifications

The timing of input and output events is an essential element of reactive systems. A linear temporal logic (LTL) is a suitable language for describing the timing of events. In this paper, we use LTL to describe the specifications of reactive systems. We treat input events and output events as atomic propositions.

Syntax. Formulae in LTL are inductively defined as follows:

- Atomic propositions (i.e., input events and output events) are formulae.
- $f \wedge g$, $\neg f$, $\mathbf{X}f$, $f\mathbf{U}g$ are formulae if f and g are formulae.

The notation $\mathbf{X}f$ means that 'f holds the next time,' while $f\mathbf{U}g$ means that 'f always holds until g holds.' The notations $f \vee g$, $f \rightarrow g$, $\top$, $f\mathbf{R}g$, $\mathbf{F}f$, and $\mathbf{G}f$ are abbreviations for $\neg(\neg f \wedge \neg g)$, $\neg(f \wedge \neg g)$, $\neg\bot$, $\neg(\neg f\mathbf{U}\neg g)$, and $\top\mathbf{U}f$, $\neg\mathbf{F}\neg f$ respectively, where $\bot$ is an atomic proposition representing 'falsity.'

Semantics. A behavior is an infinite sequence of sets of events. Let i be an index such that $i \geq 0$. The i-th set of a behavior σ is denoted by $\sigma[i]$. The i-th suffix of a behavior σ is denoted by $\sigma[i\ldots]$. When a behavior σ satisfies a formula f, we write $\sigma \models f$, and inductively define this relation as follows:

- $\sigma \models p$ iff $p \in \sigma[i]$
- $\sigma \models \neg f$ iff $\sigma \not\models f$
- $\sigma \models f \wedge g$ iff $\sigma \models f$ and $\sigma \models g$
- $\sigma \models \mathbf{X}f$ iff $\sigma[1\ldots] \models f$
- $\sigma \models f\mathbf{U}g$ iff $\exists j \geq 0.((\sigma[j\ldots] \models g)$ and $\forall k(0 \leq k < j.\ \sigma[k\ldots] \models f))$

We say that f is satisfiable if there exists a σ that satisfies f.

2.3 Properties of Reactive System Specifications

It is important for reactive system specifications to satisfy realizability. Realizability requires the existence of a reactive system such that for any input events with any timing, the system produces output events such that the specification holds.

Definition 2 (Realizability). *A specification Spec is realizable if the following holds:*

$$\exists RS \forall \tilde{a}(behave_{RS}(\tilde{a}) \models Spec),$$

where $\tilde{a}$ is an infinite sequence of sets of input events, i.e., $\tilde{a} \in (2^X)^\omega$. $behave_{RS}(\tilde{a})$ is the infinite behavior of $\tilde{a}$ caused by RS, defined as follows. If $\tilde{a} = a_0 a_1 \ldots$, $behave_{RS}(\tilde{a}) = (a_0 \cup b_0)(a_1 \cup b_1)\ldots$, where b_i is a set of output events caused by RS, i.e., $b_i = r(a_0 \ldots a_i)$.

The following property was shown to be a necessary condition for realizability in [10].

Definition 3 (Strong satisfiability). *A specification Spec is strongly satisfiable if the following holds:*

$$\forall \tilde{a} \exists \tilde{b} (\langle \tilde{a}, \tilde{b} \rangle \models \mathit{Spec}),$$

where $\tilde{b}$ is an infinite sequence of sets of output events, i.e., $\tilde{b} \in (2^Y)^\omega$. If $\tilde{a} = a_0 a_1 \ldots$ and $\tilde{b} = b_0 b_1 \ldots$, then $\langle \tilde{a}, \tilde{b} \rangle$ is defined by $\langle \tilde{a}, \tilde{b} \rangle = (a_0 \cup b_0)(a_1 \cup b_1) \ldots$.

Intuitively speaking, strong satisfiability is the property that if a reactive system is given an infinite sequence of sets of future input events, the system can determine an infinite sequence of sets of future output events. Strong satisfiability is a necessary condition for realizability; i.e., all realizable specifications are strongly satisfiable. Conversely, many practical strongly satisfiable specifications are also realizable.

Example 1. The following is a specification of a control system for a door.

1. The door has two buttons: an "open" button and a "close" button.
2. If the "open" button is pushed, the door eventually opens.
3. While the "close" button is being pushed, the door remains shut.

The events 'the "open" button is pushed' and 'the "close" button is pushed' are both input events. We denote these events by x_1 and x_2, respectively. The event 'the door is open (closed)' is an output event. We denote this event by y (resp., $\neg y$). This specification is then represented by $\mathit{Spec}_1 : \mathbf{G}((x_1 \rightarrow \mathbf{F}y) \wedge (x_2 \rightarrow \neg y))$ in LTL. This is not strongly satisfiable, and is consequently unrealizable, due to the fact that there is no response that satisfies Spec_1 for the environmental behavior in which the "close" button is still being pushed after the "open" button has been pushed. Formally, for $\tilde{a} = \{x_1, x_2\}\{x_2\}\{x_2\} \ldots$, $\exists \tilde{b} (\langle \tilde{a}, \tilde{b} \rangle \models \mathit{Spec}_1)$ does not hold. Hence $\forall \tilde{a} \exists \tilde{b} (\langle \tilde{a}, \tilde{b} \rangle \models \mathit{Spec}_1)$ does not hold.

3 Bounded Strong Satisfiability

In this section, we present the notion of bounded strong satisfiability. This property is a restricted version of strong satisfiability, in which only input sequences represented by loop structures of size k are considered.

Definition 4 (k-loop). *Let $k, l \in \mathbb{N}$ and $l \leq k$. An infinite sequence σ is a (k, l)-loop if there exists $u = s_0 s_1 \ldots s_{l-1}$ and $v = s_l s_{l+1} \ldots s_k$ such that $\sigma = u \cdot v^\omega$. An infinite sequence σ is a k-loop if there exists an l such that σ is a (k, l)-loop.*

Definition 5 (Bounded strong satisfiability). *Let $k \in \mathbb{N}$. A specification Spec is k-strongly satisfiable if the following holds:*

$$\forall \tilde{a}. (\tilde{a} \textit{ is } k\textit{-loop} \implies \exists \tilde{b}. (\langle \tilde{a}, \tilde{b} \rangle \models \mathit{Spec})).$$

If an infinite sequence is a k-loop and $k < k'$, then the sequence is a k'-loop. Therefore, the following holds:

Theorem 1. *Let $k < k'$. If a specification Spec is k'-strongly satisfiable, then Spec is also k-strongly satisfiable.*

It is clear from the definition that if a specification *Spec* is strongly satisfiable, then *Spec* is also k-strongly satisfiable. Moreover, if *Spec* is described in LTL and the bound k is sufficiently large, then the converse is also true[1].

Theorem 2. *For all $k \in \mathbb{N}$, if a specification Spec is strongly satisfiable, then Spec is also k-strongly satisfiable. Moreover, if a specification Spec is not strongly satisfiable, then Spec is not k-strongly satisfiable for some k.*

Example 2. The specification $Spec_1$ in Example 1 is not 1-strongly satisfiable, and consequently not k-strongly satisfiable for all $k > 1$, because $\tilde{a} = \{x_1, x_2\}$ $\{x_2\}$ $\{x_2\} \ldots$ is an 1-loop$(\{x_1, x_2\}\{x_2\}^\omega)$, which does not satisfy $\exists \tilde{b}(\langle \tilde{a}, \tilde{b} \rangle \models Spec_1)$.

By checking whether or not a specification satisfies bounded strong satisfiability, we can know whether or not the specification has simple input patterns (represented by small loops) that cannot satisfy the specification. As the experiment in Section 6 shows, many practical defective specifications have such simple input patterns. Thus, we anticipate that checking bounded strong satisfiability will find many practical defects in reactive system specifications.

4 Procedure for Checking Bounded Strong Satisfiability

In this section, we present a procedure for checking bounded strong satisfiability using non-deterministic Büchi automata. This procedure is based on the procedure for (unbounded) strong satisfiability introduced in [6].

A *non-deterministic Büchi automaton*(NBA) is a tuple $\mathcal{A} = \langle \Sigma, Q, q_0, \delta, F \rangle$, where Σ is an alphabet, Q is a finite set of states, q_0 is an initial state, $\delta \subseteq Q \times \Sigma \times Q$ is a transition relation, and $F \subseteq Q$ is a set of final states. A run of $\mathcal{A}$ on an ω-word $\sigma = \sigma[0]\sigma[1] \ldots$ is an infinite sequence $\varrho = \varrho[0]\varrho[1] \ldots$ of states, where $\varrho[0] = q_0$ and $(\varrho[i], \sigma[i], \varrho[i+1]) \in \delta$ for all $i \geq 0$. We say that $\mathcal{A}$ accepts σ, if there is a run ϱ on σ such that $Inf(\varrho) \cap F \neq \emptyset$ holds, where $Inf(\varrho)$ is the set of states that occurs infinitely often in ϱ. The set of ω-words accepted by $\mathcal{A}$ is called the language accepted by $\mathcal{A}$, and is denoted by $L(\mathcal{A})$.

Let *Spec* be a specification written in LTL. We can check the bounded strong satisfiability of *Spec* via the following procedure.

1. We obtain an NBA $\mathcal{A} = \langle 2^{X \cup Y}, Q, q_0, \delta, F \rangle$ s.t. $L(\mathcal{A}) = \{\sigma | \sigma \models Spec\}$ holds.
2. Let $\mathcal{A}' = \langle 2^X, Q, q_0, \delta', F \rangle$ be the NBA obtained by restricting $\mathcal{A}$ to only input events, where $\delta' = \{(q, a, q') | \exists b \ (q, a \cup b, q') \in \delta\}$. Note that $L(\mathcal{A}') = \{\tilde{a} | \exists \tilde{b} \langle \tilde{a}, \tilde{b} \rangle \in L(\mathcal{A})\}$ holds due to the definition of δ'.
3. We check whether or not $\mathcal{A}'$ accepts all k-loops (*i.e., is k-universally acceptable*). If it is k-universally acceptable, we conclude that *Spec* is k-strongly satisfiable. Otherwise, we conclude that *Spec* is not k-strongly satisfiable.

The construction of the NBA in step 1 can be based on the tableau methods of [2] and others. In step 3, bounded universality is checked using an SAT solver.

[1] This is derived from the fact that *Spec* can be represented by a finite state Büchi automaton.

5 SAT-Based Bounded Universality Checking for NBA

We present an SAT-based method for checking the bounded universality of an NBA. In this method, the complement of the bounded universality checking problem is reduced to an SAT problem.

5.1 Characterization of Non-accepted k-Loops

As a preliminary to the reduction, we characterize the k-loops that are not accepted by NBA, based on the notion of a run graph.

Definition 6 (run graph). *Let $\mathcal{A} = (\Sigma, Q, q_I, \delta, F)$ be NBA and $\sigma = s_0 \dots s_{l-1} (s_l \dots s_k)^\omega$ be a (k,l)-loop. The run graph for $\mathcal{A}$ and σ is $G = (V, v_I, E, C)$: $V := Q \times \{0, 1, \dots, k\}$ (the set of nodes). $v_I := (q_I, 0)$ (the initial node). $E := \{((q,i), (q', suc(i)))|(q, s_i, q') \in \delta\}$, where $suc(i) = l$ if $i = k$, and $suc(i) = i + 1$ otherwise (the set of edges). $C := \{(q,i)|q \in F, 0 \leq i \leq k\}$ (the set of final nodes).*

An NBA does not accept a (k,l)-loop σ if and only if there does not exist a run ϱ on σ such that $Inf(\varrho) \cap F = \emptyset$ holds; i.e., for all runs, the number of occurrences of final states in the run is finite. A run on $\sigma[i \dots]$ from a state q corresponds to a path in the run graph from (q, i). Then the following holds:

Theorem 3. *An NBA $\mathcal{A}$ does not accept a (k,l)-loop σ if and only if, for all paths from the initial node in the run graph for $\mathcal{A}$ and σ, the number of occurrences of final nodes in the path is finite.*

The number of final nodes in a run graph is bounded. From this, the following result is derived [2]:

Lemma 1. *Let $d \geq |C|$. If for all paths $\tilde{v}$ in a run graph G from a state v, the number of occurrences of final nodes in $\tilde{v}$ is finite, then for all paths $\tilde{v}$ from v in G, final nodes occur at most d times in $\tilde{v}$.*

The property that "for all paths $\tilde{v}$ from v, the number of occurrences of final nodes is at most j"(denoted by $AtMost(v, j)$) is characterized as follows: For $v \in V \setminus C$ (w.r.t. $v \in C$), $AtMost(v, j)$ holds if and only if for all successors $v' \in vE$, $AtMost(v', j)$ holds(w.r.t. $AtMost(v', j-1)$ holds). Additionally, for all $v \in C$, $AtMost(v, 0)$ does not hold. Based on this idea, the following result can be proved:

Theorem 4. *Let $G = (V, v_I, E, C)$ be a run graph and $d \in \mathbb{N}$. For all paths $\tilde{v}$ from v_I in G, final nodes occur at most d times if and only if there exist sets of nodes $V_0, V_1, \dots, V_d$ such that the following are true:*

[2] If the number of occurrences of final nodes in a path is more than $|C|$, then there exists a final state q_c that occurs at least twice, which implies the existence of a path on which the final state q_c occurs infinitely often.

1. *The following condition (denoted by $I(V_0)$) holds:*

$$v \in V_0 \iff \begin{cases} \forall v' \in vE.\ v' \in V_0 & \text{if } v \in V \setminus C, \\ \bot & \text{if } v \in C \end{cases}$$

2. *For all $0 \leq j < d$, the following condition (denoted by $T(V_j, V_{j+1})$) holds:*

$$v \in V_{j+1} \iff \begin{cases} \forall v' \in vE.\ v' \in V_{j+1} & \text{if } v \in V \setminus C \\ \forall v' \in vE.\ v' \in V_j & \text{if } v \in C \end{cases}$$

3. *$v_I \in V_d$ holds (denoted by $F(V_d)$).*

We can summarize the characterization of non-accepted k-loops in the following result:

Theorem 5. *Let $\mathcal{A} = (\Sigma, Q, q_I, \delta, F)$ be an NBA and $k \in \mathbb{N}$. For all $d \in \mathbb{N}$, (2) implies (1), and for $d \geq (k+1) \cdot |F|$, (1) implies (2), where (1) and (2) are as follows:*

(1) There exists a k-loop that is not accepted by $\mathcal{A}$.
(2) There exists a k-loop σ such that for some sets $V_0, V_1, \ldots, V_d$ of nodes of the run graph G for $\mathcal{A}$ and σ, $I(V_0) \wedge \bigwedge_{0 \leq j < d} (V_j, V_{j+1}) \wedge F(V_d)$ holds.

5.2 Reduction to SAT

We present a reduction to SAT based on Theorem 5. In other words, for an NBA $\mathcal{A}$ and k, we construct a propositional formula $|[notAcc(\mathcal{A}, k, d)]|$ such that condition (2) of Theorem 5 holds if and only if $|[notAcc(\mathcal{A}, k, d)]|$ is satisfiable.

Variables. To represent a (k, l)-loop and $V_0, V_1, \ldots, V_d$, we introduce the following variables (assuming that $\Sigma = 2^P$): (a) p_i for each $p \in P$, $0 \leq i \leq k$, which indicate whether or not the i-th element s_i of a k-loop satisfies $p \in s_i$; (b) l_i for $0 \leq i \leq k$, which indicate whether or not the i-th element follows the k-th element; (c) $v^j_{(q,i)}$ for $q \in Q, 0 \leq i \leq k, 0 \leq j \leq d$, which indicate whether or not $(q, i) \in V_j$ holds.

Constraint. To represent the concept "be a k-loop correctly", we define the formula $|[loop(k)]| := \bigvee_{0 \leq i \leq k} l_i \wedge \bigwedge_{0 \leq i \leq k} (l_i \to \bigwedge_{0 \leq i' \leq k, i' \neq i} \neg l_{i'})$.

To represent $I_\sigma \wedge \bigwedge_{0 \leq j < d} T(V_j, V_{j+1}) \wedge F(V_d)$, we define the formulae $|[I(\mathcal{A}, k)]|_0$ $|[T(\mathcal{A}, k)]|_{j,j+1}$ and $|[F(\mathcal{A}, k)]|_d$, which respectively indicate that $I(V_0)$, $T(V_i, V_{i+1})$ and $F(V_d)$ hold. Let $\mathcal{A} = (2^P, Q, q_I, \delta, F)$ be an NBA and $k \in \mathbb{N}$. These formulae are defined in Table 1, where $|[a]|_i$ is the formula indicating that the i-th element of σ is a, and $|[suc]|^j_{(q,i)}$ is the formula indicating that $(q, suc(i)) \in V_j$ holds.

The formula $|[notAcc(\mathcal{A}, k, d)]|$ is defined by $|[notAcc(\mathcal{A}, k, d)]| := |[loop(k)]| \wedge |[I(\mathcal{A}, k)]|_0 \wedge \bigwedge_{0 \leq i < d} |[T(\mathcal{A}, k)]|_{j,j+1} \wedge |[F(\mathcal{A}, k)]|_d$.

Table 1. The definitions of $|[I(\mathcal{A},k)]|_0$, $|[T(\mathcal{A},k)]|_{j,j+1}$ and $|[T(\mathcal{A},k)]|_{j,j+1}$

$\|[I(\mathcal{A},k)]\|_0$	$\bigwedge_{\substack{q\in Q\setminus F,\\ 0\le i\le k}} \left(v^0_{(q,i)} \leftrightarrow \bigwedge_{(q,a,q')\in\delta} (\|[a]\|_i \rightarrow \|[suc]\|^0_{(q',i)}) \right) \wedge \bigwedge_{\substack{q\in F,\\ 0\le i\le k}} \left(\neg v^0_{(q,i)}\right)$
$\|[T(\mathcal{A},k)]\|_{j,j+1}$	$\bigwedge_{\substack{q\in Q\setminus F,\\ 0\le i\le k}} \left(v^{j+1}_{(q,i)} \leftrightarrow \bigwedge_{(q,a,q')\in\delta} (\|[a]\|_i \rightarrow \|[suc]\|^{j+1}_{(q',i)}) \right) \wedge \bigwedge_{\substack{q\in F,\\ 0\le i\le k}} \left(v^{j+1}_{(q,i)} \leftrightarrow \bigwedge_{(q,a,q')\in\delta} (\|[a]\|_i \rightarrow \|[suc]\|^{j}_{(q',i)}) \right)$
$\|[F(\mathcal{A},k)]\|_d$	$v^d_{(q_I,0)}$

Theorem 6. *Let $\mathcal{A}$ be an NBA and $k \in \mathbb{N}$. For all $d \in \mathbb{N}$, (2) implies (1), and for $d \geq (k+1)\cdot|F|$, (1) implies (2), where (1) and (2) are as follows:*

1. *There exists a k-loop which is not accepted by $\mathcal{A}$.*
2. *$|[notAcc(\mathcal{A},k,d)]|$ is satisfiable.*

Theorem 7. *Let $\mathcal{A} = (\Sigma, Q, q_I, \delta, F)$ and $k, d \in \mathbb{N}$. The size of $|[notAcc(\mathcal{A},k,d)]|$ is $\mathcal{O}(k^2 + k\cdot d\cdot|\delta|)$. Checking that $\mathcal{A}$ is not k-universally acceptable can be reduced to the SAT problem for a formula of size $\mathcal{O}(k^2\cdot|F|\cdot|\delta|)$.*

5.3 Improvement

Incremental Checking. The condition of Theorem 5 can be regarded as the reachability problem of a transition system for which the initial condition is I, the transition relation is T, and the final condition is F. Consequently, the incremental technique of [14] can be applied to our method. Using this technique, k-universality checking can be accomplished for small d.

Reduction Based on a Modified Run Graph. Checking that an NBA is not k-universally acceptable can be also reduced to the SAT problem for a formula of size $\mathcal{O}(k\cdot|Q|\cdot|\delta|)$ by modifying the construction of the run graph as follows. We add a check bit to each node: $V' := Q\times\{0,1,\ldots,k\}\times\{\top,\bot\}$. The check bit indicates whether or not final nodes occurred before the given node was reached in each turn. We also define $C' := \{(q,k,\top)|q\in Q\}$. The modified run graph has the same features as the normal run graph. The size $|V'|$ is also $\mathcal{O}(k\cdot|Q|)$, but the size $|C'|$ is $|Q|$, while the size $|C|$ is $(k+1)\cdot|F|$. Thus $\mathcal{O}(k\cdot|Q|\cdot|\delta|)$-encoding is possible.

6 Experiments

We implemented our method and compared its execution time to that of realizability and (unbounded) strong satisfiability[3].

Our implementation (denoted by BSS) is as follows. Steps 1 and 2 in the procedure of Section 4 are based on [2]. The k-universality checking of Step 3

[3] All experiments were performed on a Pentium(R) D 3.0 GHz with 2.0 GB RAM.

Table 2. The checking times (sec) for each property

$Spec_n$	BSS $k=1$	SS	R Lily
2	0.05	219.32	955.23
3	0.13	>3600	>3600
4	1.17		
5	12.95		
6	181.62		
7	1962.62		

$Spec'_n$	BSS $k=1$	BSS $k=5$	BSS $k=10$	SS	R lily
2	0.06	0.12	0.21	>3600	>3600
3	0.17	1.12	2.00		
4	1.60	44.08	82.95		
5	24.76	>3600	>3600		
6	297.59				
7	2942.89				

is accomplished incrementally for $d = 0, 1, \ldots$, based on the modified run graph (described in Section 5.3). We use MiniSat[5] as the SAT solver. To check (unbounded) strong satisfiability (denoted by SS), we check (unbounded) universality. Universality is checked with the symbolic model checker NuSMV[4]. To check realizability (denoted by R), we use Lily[9].

Table 2 lists the checking times for the specification $Spec_n$ in [2] (a specification for an elevator control system involving n floors) and $Spec'_n$, which includes a fairness assumption. The specification $Spec_n$ and $Spec'_n$ have $3n + 6$ atomic propostions ($|X| = n + 2$, $|Y| = 2n + 4$). The numbers of the occurrence of temporal operators in $Spec_n$ and $Spec'_n$ are $6n - 1$ and $7n$, respectively. In all tests of $Spec_n$, the judgment was "NO." The results indicate that our method can handle larger specifications more efficiently, and can obtain the simple input pattern of a defect in $Spec_n$. Because bounded strong satisfiability is a necessary condition for realizability, our method also successfully showed that $Spec_n$ is not unrealizable for $n > 2$, which the realizability checker failed to indicate. In all tests of $Spec'_n$, the judgment was "Yes.' Hence, our method showed that for any simple input pattern, there is a response that satisfies $Spec'_n$ in real time.

7 Related Work

Encoding Methods for Bounded Model Checking. For bounded model checking, SAT encoding methods of problems whether or not models satisfy specifications represented by LTL, very weak alternating Büchi automata (VWABA) and weak alternating Büchi automata (WABA) are presented in [3,15,7]. LTL and VWABA are less expressive than NBA, which was used in this work. WABA and NBA are of equal expressiveness. Indeed, bounded universality checking for NBA can also be accomplished via WABA, using the WABA encoding method of [7]. However, our method is more efficient than the WABA approach[4].

Bounded Realizability. In [13], the notion of bounded realizability and a checking method using an SMT solver are presented. Bounded realizability is the property that there exists a reactive system that acts as a transition system of k states, such that all behaviors satisfy the specification. In bounded realizability checking, transition systems of k states are searched. On the other hand, in our method, k-loops (which are simpler) are searched. Because of this, our method can detect simple and typical defects in larger specifications.

[4] The total size of the propositional formulae of the WABA approach is $\mathcal{O}(k \cdot |Q|^2 \cdot |\delta|)$, whereas we have provided $\mathcal{O}(k \cdot |Q| \cdot |\delta|)$-encoding.

8 Conclusion

We introduced the notion of bounded strong satisfiability and a checking method for this property, for detecting simple defects in reactive system specifications. In our method, we construct an NBA that accepts input sequences for which there is no response that satisfies a specification, then check whether or not the NBA is k-universally acceptable, using an SAT solver. We implemented our method and demonstrated that it can handle larger specifications than the checking techniques for other properties, and can also detect simple defects efficiently.

References

1. Abadi, M., Lamport, L., Wolper, P.: Realizable and Unrealizable Specifications of Reactive Systems. In: Ronchi Della Rocca, S., Ausiello, G., Dezani-Ciancaglini, M. (eds.) ICALP 1989. LNCS, vol. 372, pp. 1–17. Springer, Heidelberg (1989)
2. Aoshima, T., Sakuma, K., Yonezaki, N.: An efficient verification procedure supporting evolution of reactive system specifications. In: Proc. International Workshop on Principles of Software Evolution, pp. 182–185 (2001)
3. Biere, A., Cimatti, A., Clarke, E., Zhu, Y.: Symbolic Model Checking without BDDs. In: Cleaveland, W.R. (ed.) TACAS 1999. LNCS, vol. 1579, pp. 193–207. Springer, Heidelberg (1999)
4. Cimatti, A., Clarke, E., Giunchiglia, E., Giunchiglia, F., Pistore, M., Roveri, M., Sebastiani, R., Tacchella, A.: NuSMV 2: An OpenSource Tool for Symbolic Model Checking. In: Brinksma, E., Larsen, K.G. (eds.) CAV 2002. LNCS, vol. 2404, pp. 359–364. Springer, Heidelberg (2002)
5. Eén, N., Sörensson, N.: An Extensible SAT-solver. In: Giunchiglia, E., Tacchella, A. (eds.) SAT 2003. LNCS, vol. 2919, pp. 502–518. Springer, Heidelberg (2004)
6. Hagihara, S., Yonezaki, N.: Completeness of verification methods for approaching to realizable reactive specifications. In: Proc. Asian Working Conference on Verified Software, pp. 242–257 (2006)
7. Heljanko, K., Junttila, T.A., Keinänen, M., Lange, M., Latvala, T.: Bounded Model Checking for Weak Alternating Büchi Automata. In: Ball, T., Jones, R.B. (eds.) CAV 2006. LNCS, vol. 4144, pp. 95–108. Springer, Heidelberg (2006)
8. Jackson, D.: Software Abstractions: Logic, Language, and Analysis. The MIT Press (2006)
9. Jobstmann, B., Bloem, R.: Optimizations for LTL synthesis. In: Proc. FMCAD, pp. 117–124 (2006)
10. Mori, R., Yonezaki, N.: Several realizability concepts in reactive objects. In: Proc. Information Modeling and Knowledge Bases IV: Concepts, Methods and Systems, pp. 407–424. IOS Press (1993)
11. Mori, R., Yonezaki, N.: Derivation of the Input Conditional Formula from a Reactive System Specification in Temporal Logic. In: Langmaack, H., de Roever, W.-P., Vytopil, J. (eds.) FTRTFT 1994 and ProCoS 1994. LNCS, vol. 863, pp. 567–582. Springer, Heidelberg (1994)
12. Pnueli, A., Rosner, R.: On the synthesis of a reactive module. In: Proc. POPL, pp. 179–190 (1989)
13. Schewe, S., Finkbeiner, B.: Bounded Synthesis. In: Namjoshi, K.S., Yoneda, T., Higashino, T., Okamura, Y. (eds.) ATVA 2007. LNCS, vol. 4762, pp. 474–488. Springer, Heidelberg (2007)

14. Sheeran, M., Singh, S., Stålmarck, G.: Checking Safety Properties Using Induction and a SAT-Solver. In: Johnson, S.D., Hunt Jr., W.A. (eds.) FMCAD 2000. LNCS, vol. 1954, pp. 108–125. Springer, Heidelberg (2000)
15. Sheridan, D.: Bounded model checking with SNF, alternating automata, and Büchi automata. Electron. Notes Theor. Comput. Sci. 119(2), 83–101 (2005)
16. Shimakawa, M., Hagihara, S., Yonezaki, N.: Complexity of checking strong satisfiability of reactive system specifications. In: Proc. International Conference on Advances in Information Technology and Communication, pp. 42–51 (2012)

OSMF: A Framework for OSS Process Measurement

Wikan Danar Sunindyo and Fajar Juang Ekaputra

Data and Software Engineering Research Group
School of Electrical Engineering and Informatics, Bandung Institute of Technology
Labtek V 2nd floor Ganesha Street 10 Bandung 40132 Indonesia
{wikan,fajar}@informatika.org

Abstract. An Open Source Software (OSS) project can be considered as a new type of business entity involving various roles and stakeholders, e.g., project managers, developers, and users, who apply individual methods. The project managers have the responsibility to manage the OSS development in a way that the OSS product can be delivered to the customers in time and with good quality. This responsibility is challenging, because the heterogeneity of the data collected and analyzed from different stakeholders leads to the complexity of efforts of the project managers to measure and manage OSS projects. In this paper, we propose a measurement framework (OSMF) to enable the project managers to collect and analyze process data from OSS projects efficiently. Initial results show that OSMF can help project managers to manage OSS business processes more efficient, hence improve the decision on OSS project quality.

Keywords: Open Source Software, process observation, project management, software quality.

1 Introduction

Both development and usage of Open Source Software (OSS) products are increasing exponentially and play a significant role in the software economy [4]. OSS development projects are business entities, which involve several different stakeholders, like project managers, developers, and customers [13].

Mockus *et al.* [8] describe four key characteristics of OSS projects: first, it consist of large number of volunteers [9]; second, anyone can choose tasks he/she wants to execute [3]; third, there is neither explicit system design nor explicit detailed design [5]; fourth, there is no project plan, schedule or list of deliverables [8].

Major OSS projects like FreeBSD divide project participants into three main roles: *core team members*, *committers* and *contributors* [14]. The Core Team is a small group of senior developers that act as project managers who are responsible for deciding about the overall goals and directions of the project. Committers are developers who have the authority to commit changes to the project repository. Contributors are people who want to contribute to the project, but do not have committer privileges.

Typically, project managers face challenges defining, manage and improve business process in OSS projects, since business processes in OSS projects are different than business processes in conventional software projects. However, business process

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 71–80, 2013.

definition is still a key foundation to enable and improve business process observation by means of monitoring and controlling the status of OSS projects.

In this paper, we propose a measurement framework (OSS project measurement framework – OSMF) to manage (i.e., definition, observation, and control) business processes in OSS projects. The contribution of this framework is to guide the project manager in following the processes of engineering process observation.

For evaluating the OSMF, we use OSS project data from the Red Hat Enterprise Linux (RHEL) and Fedora projects as use cases. We evaluate the engineering processes in both projects, by checking conformance between the engineering process models from both projects to the designed process model. The results can show the frequent states of the process for project managers' decision.

We organize this paper as follows. After introduction, we explain related works that support the OSMF framework. The third section presents the research issues that we discuss in this paper. Section four shows the case we use in our work. Section five discusses about the results of our study. We summarize the results and lessons learned of our work in section six, and finally conclude our paper and present future work.

2 Related Work

This section summarizes related works on business process management and OSS.

2.1 Business Process Management

Business process management (BPM) is a management approach focusing on aligning all aspects of an organization with the requests and needs of clients. It is a holistic management approach [16] that promotes business effectiveness and efficiency while striving for innovation, flexibility, and integration with technology. An empirical study by Kohlbacher reveals that BPM helps organizations to gain higher customer satisfaction, product quality, delivery speed and time-to-market speed [6].

Reijers, van der Aalst and zur Muehlen used the notion of a life-cycle to distinguish between the various phases that a BPM initiative can go through [11, 15, 17].

This life-cycle distinguishes between six phases: namely analysis, design, implementation, enactment, monitoring, and evaluation phases. In the *analysis* phase, a set of requirements is developed for the business process in questions such as performance goals or intentions. In the *design* phase, the process activities, their order, the assignment of resources to activities and the organization structure are defined. In the *implementation* phase, the infrastructure for the business process is set up. The dedicated infrastructure is used to handle individual cases covered in the *enactment* phase. In the *monitoring* phase, counteractions are taken to deal with problematic situations depending on process metrics. The new requirements are taken as input in the next turn of the business process management life-cycle in the *evaluation* phase [10].

2.2 Process Mining

Process mining is a process management technique that allow for the analysis of business processes based on event logs. Process mining aims at improving this by providing techniques and tools for discovering process, control, data, organizational, and social structures from event logs[1].

In the area of software development project, Rubin et al. [12] proposed a process mining framework for software processes, which is based on fact that software development process are often not explicitly modeled and even chaotic. In this paper, we use this approach to develop our own framework for mining OSS processes.

3 Research Issues

The major challenges related to managing business processes in OSS projects are as follows, (1) better definition of OSS business processes, (2) more effective data collection for heterogeneous OSS business processes, (3) better analysis methods to give more precise results for project managers' decision making. From these challenges, we derive the research issues addressed in this paper as follows:

RI-1: How to describe the business aspects and their interactions in OSS projects. OSS project is a unique business entity, because it mixes the conventional business process and non-conventional business process.

We propose to describe business aspects in OSS projects and their interactions between business managers, project developers, and customers, so we get clearer view on the OSS projects business.

RI-2: How to evaluate measurement of the engineering processes data? The measurement of the OSS projects engineering process data can be done by collecting and analyzing engineering process data from OSS projects.

We propose to measure the bug history data by performing conformance check analyses between the actual process model and the designed process model. We defined two research hypotheses to be validated by the experiments as follows.

RH-1. RHEL and Fedora developers are following the naming and order of the bug states of the original Bugzilla life cycle. The changing of bug states from developers lead to a chain of bug status history that we can trace to find out the pattern of the bug status usually used by the developers in developing the OSS projects.

In this study, we generate process models from bug history data of two OSS projects, namely RHEL and Fedora and make conformance checking with the designed process model from Bugzilla. IF ϕ is the designed process model, and ψ is an OSS project, and P is a function to get the number of bug states from designed process model or from OSS projects, THEN we can formulate following null hypothesis.

$$\boldsymbol{H01}: \{\exists \psi \in (\text{RHEL, Fedora}) \mid P(\psi) = P(\phi)\} \quad \text{(eq. 1)}$$

[1] http://www.processmining.org

RH-2. OSS projects developers are using all bug states for each bug history in the same number of frequency. We want to measure and investigate the frequency of bug states used in one OSS project. IF s_i is a bug state and s_{i+1} is another bug state after s_i, and N is a function to obtain the frequency of bug states using in Fedora, THEN we can propose following null hypothesis.

$$\boldsymbol{H02}: \{\forall s_i,\ s_{i+1} \mid N(s_i) = N(s_{i+1})\} \quad \text{(eq. 2)}$$

4 Solution Approach

In this section we address the research issues in section 3.

4.1 Business Aspects of OSS Projects

Crowston et al. [2] suggested three aspects of the success of OSS projects, namely *developers' contribution*, the intensity of *software usage*, and the *quality of software products*. Based on this suggestion, we model the business aspects of OSS projects in three parts, namely *developer*, *customer*, and *business* parts. The causality model of the business aspects and their interactions is shown in Figure 2.

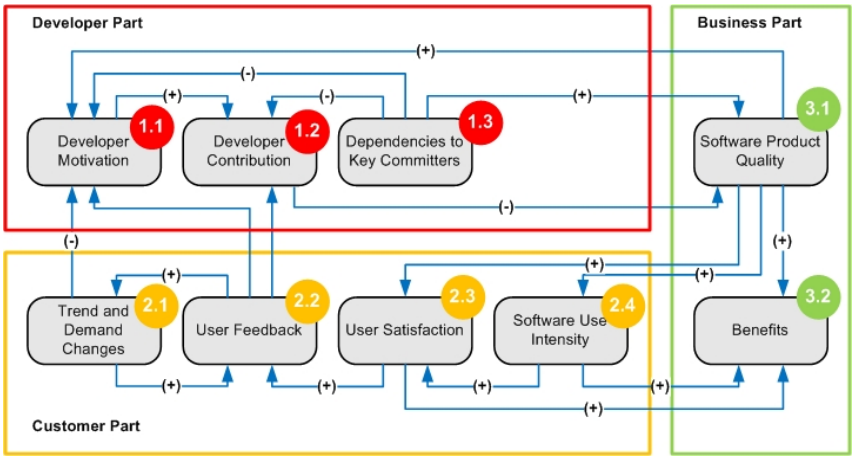


Fig. 1. Model of OSS Projects Business Aspects and Their Interactions

The developer part consists of three aspects. (1.1) *Developer motivation* is about the motivation of developers to join the developers' community to develop the product. (1.2) *Developer contribution* consists of contribution of developers to the OSS project, e.g., via source code management, developers' mailing list, or bug management systems. (1.3) *Dependencies to key committers* mean that the survivability of the OSS projects depends on very few active committers (key committers).

The customer part consists of four aspects. (2.1) *Trend and Demand Changes* from the customers can be submitted to the developers and be a part of software

requirements. (2.2) *User feedback* is a feedback from the users about functionality of the software that can be improved by the developers. (2.3) *User satisfaction* is a notification from the users that they satisfy with the functionality of the product. (2.4) *Software use intensity* is a intensity of software usage by the users.

The business part deals with software product quality and benefits of the software products. (3.1) a *software product quality* deals with the qualitative and quantitative measurement of product that comply with users' satisfaction. (3.2) the *benefits* of software product can be used as a selling point of the product.

4.2 OSMF Measurement Framework

To analyze and measure the OSS processes, we propose the OSMF measurement framework for supporting the OSS project managers with data collection and data analysis. Figure 3 shows the measurement framework for observing OSS processes, i.e. bug history data.

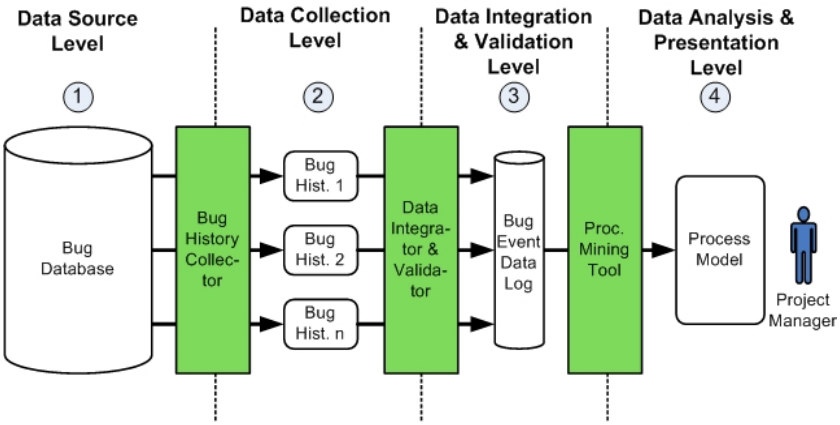


Fig. 2. OSMF Framework for OSS Processes Measurement

This framework consists of 4 levels, namely data source level, data collection level, data integration and validation level, and data analysis and presentation level.

(1) In the *data source level*, we have bug database which contains all bugs information that are used in software development.
(2) In *data collection level,* we extract and collect bug history data from bug history data using a bug history collector. The bug history collector is written in Java and use XML RPC[2] web service interface to access to the bug data.
(3) *Data integration and validation.* The collected bug history data is integrated and validated using data integrator and validator. The integration and validation process is implemented by using Engineering Knowledge Base (EKB) [1].

[2] `https://bugzilla.redhat.com/xmlrpc.cgi`

(4) *Data analysis and presentation*. The even data log from previous level is analyzed by using Process Mining tool. The results are presented to the project managers.

The study objects of this research are two different OSS projects, namely the Red Hat Enterprise Linux and Fedora projects. Both projects are developing Linux-based operating systems and under the same Red Hat project management.

Currently, in the Red Hat Bugzilla browser, there are in total 23.768 bugs reported RHEL version 6 (7.465 open bugs and 16.303 closed bugs) and 293.685 bugs reported Fedora (25.386 open bugs and 268.299 closed bugs).

4.3 Evaluation of Engineering Process Measurement Results

For analyzing the bug history data, we used a process analysis tool called ProM. This tool has capabilities to discover process model, make conformance checking between expected process model and the process model generated from actual data, and make performance analysis on process model for process improvement. There are a lot of plug-ins and algorithms to discover the process model from actual data. One of them is heuristics mining.

The heuristics mining is a process mining algorithm in ProM which is based on the frequency of the patterns. The most important characteristic of the heuristics mining is the robustness for noise and exceptions. We use the heuristics mining to analyze event log from bug history data to find out the process model from actual data, rather than designed process model.

We identified and addressed threats to internal and external validity of our evaluation results as follows.

Threats to Internal Validity. As we have conducted previous experiments using fewer data, there is a tendency of increasing of the numbers of states as new data is added. So we put more focus on the frequency of states taken during development, since the number of states can be unnecessary increasing, while the top states remain stable.

Threats to External Validity. In this study we focus on projects with similar characteristics that may raise concerns whether the results are also valid for other project contexts. While we assume our approach to hold for projects similar to our objects, further research work is necessary to investigate projects with strongly differing characteristics.

5 Results

In this section, we report the results of our study presented in section 4.

5.1 OSS Processes Measurement Framework

To observe software engineering processes from the bug database effectively, we applied the observation framework from figure 3. In the application, we take Bugzilla

report of RHEL and Fedora projects as data sources, Bug History Data Collector as an automated data collector, integrator and validator, and Process Mining (ProM) tool for data analysis and presentation.

We have implemented and used a Java-based Bug History Data Collector tool to collect, integrate, and validate bug history data from Bugzilla database. This tool can select bug ids based on the OSS projects and versions. As results, we have collected 1000 data sets from RHEL 6 and Fedora 14. These data will be analyzed for process model conformance checking with the Bugzilla life cycle.

5.2 Evaluation of Engineering Process Measurement Results

We analyze the number of states in the process models generated by ProM and count the frequency of each state for each RHEL version. We compare the results with the designed process model from Bugzilla life cycle and evaluate the actual data by answering the two hypotheses defined in section 3.

Table 1. Name and frequency of Bug States used in RHEL and Fedora 14, compared with designed process model from Bugzilla life cycle

States	Bugzilla LC	RHEL 6		Fedora 14	
		Occ. (abs)	Occ. (rel)	Occ. (abs)	Occ. (rel)
CLOSED	✓	**546**	**33.6 %**	**500**	**54.1 %**
ASSIGNED	✓	**255**	**15.7 %**	**208**	**22.5 %**
NEEDINFO	×	6	0.37 %	**73**	**7.9 %**
MODIFIED	×	**306**	**18.9 %**	**74**	**8.0 %**
REOPENED	✓	1	0.1 %	×	×
ON_QA	×	**259**	**16.0 %**	**43**	**4.7 %**
RELEASE_PENDING	×	×	×	1	0.1 %
NEW	✓	7	0.4 %	19	2.1 %
NEEDINFO_REPORTER	×	2	0.1 %	1	0.1 %
INVESTIGATE	×	2	0.1 %	×	×
VERIFIED	✓	**199**	**12.3 %**	2	0.2 %
FAILS_QA	×	×	×	1	0.1 %
ASSIGN_TO_PM	×	1	0.1 %	×	×
ON_DEV	×	8	0.5 %	2	0.2 %
POST	×	31	1.9 %	×	×
UNCONFIRMED	✓	×	×	×	×
RESOLVED	✓	×	×	×	×

RHEL and Fedora developers are following the naming and order of the bug states of the original Bugzilla life cycle. Table 1 shows the comparison of bug states used in the Bugzilla life cycle, RHEL 6, and Fedora 14. From Table 1 we can see different bug state names are used during addressing bug in different RHEL versions.

RHEL and Fedora developers are using all bugs states for each bug history in the same frequency. Table 1 shows the frequencies of the using of each bug state in the bug history. We can see that the frequencies for different bugs in one RHEL version are not similar. The usage of some states is more frequent then of other states.

6 Discussion

In this section, we discuss our results based on the research issues.

6.1 OSS Processes Measurement Framework

We followed the observation framework we defined earlier to improve the engineering process in OSS projects. The benefit of this framework is an effective and systematic approach to collect and analyze data from the bug database to support the project managers' decision making to improve the process quality in OSS projects.

Data collection is done automatically by using our bug history data collector tool. The data integration and validation is done by using Engineering Knowledge Base (EKB). By using OSMF to collect, integrate and validate the data, we have a clean data to be analyzed using ProM. The OSMF can make analysis on the integrated data easier than analysis on single separated data.

6.2 Evaluation of Engineering Process Measurement Results

The evaluation of actual engineering process model is done by checking its conformance to the designed process model. The process model that is generated based on Fedora bug history data is shown in Figure 4(A). The designed process model (Bugzilla Life Cycle) shown in Figure 4(B).

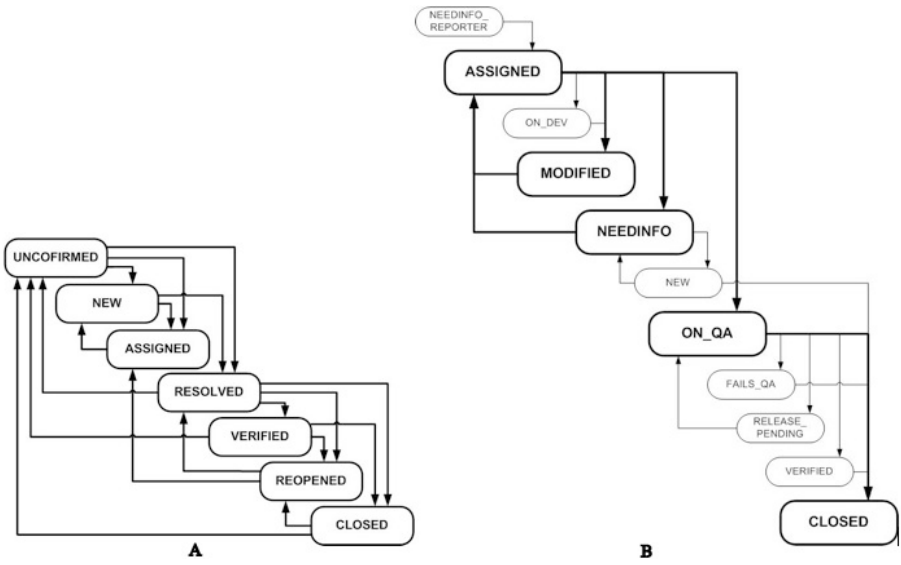


Fig. 3. Process Model of Bug Status. (A) Bugzilla LC. (B) Fedora 14

From the process model generation on RHEL and Fedora, we answer the two defined research hypotheses as follows:

RHEL and Fedora developers are following the naming and the ordering of the bug states from Bugzilla life cycle. As shown in table 1, we can see the differences of number of states used in the designed process model (Bugzilla life cycle) and states used in the generated process models from RHEL and Fedora. Therefore $\{\exists\psi \in (\text{RHEL,Fedora}) \mid P(\psi) \neq P(\phi)\}$ thus we can reject our null hypothesis H01.

An interpretation of these results can be the fact that the number of bug states available and used in RHEL and Fedora are different, meaning that both OSS projects are using different development strategy.

OSS projects developers are using all bug states for each bug history in the same number of frequency. As shown in table 2, each bug state is used in different frequency by the developers. Some bug states are used more often than the others. Therefore $\{\forall si,si+1 \mid N(si) \neq N(si+1)\}$ thus we can reject our null hypothesis H02.

An interpretation of these results can be the fact that the typically OSS projects do not follow a strict waterfall-like software engineering process, but rather a sometimes mixed dynamic software engineering process.

The proposed approach can be generalized to test other hypothesis and work on other kind of OSS project data, with some adaptations depend on the type of bug reporting tools used.

7 Summary and Future Work

Measurements of OSS processes are needed as an initial way to improve the quality of OSS processes and products. OSS project managers need to collect, integrate and analyze heterogeneous data originating from different tools used in the OSS project, such as source code management, developer's mailing list, and bug reporting tools to observe the processes and determine the status of OSS projects.

In this paper, we have explained the contribution of a measurement framework in improving the process quality in OSS projects. We used bug history data from Red Hat Enterprise Linux (RHEL) and Fedora projects as a use case for our measurement framework application and use the Heuristics Mining algorithm of the Process Mining tool ProM. The analysis results on conformance checking of process models from RHEL and Fedora bug history data can be used to improve the process quality.

Future Work. Future work will include the risk analysis on handling the bugs in the OSS project development in order to improve the quality of OSS products. Data sources that currently limited to bug report will be expanded to include other sources.

Acknowledgments. This work has been supported by the Christian Doppler Forschungsgesellschaft, the BMWFJ, Austria and the Ministry of Education, Republic of Indonesia.

References

1. Biffl, S., Sunindyo, W.D., Moser, T.: Semantic Integration of Heterogeneous Data Sources for Monitoring Frequent-Release Software Projects. In: International Conference on Complex, Intelligent and Software Intensive Systems, pp. 360–367. IEEE Computer Society (2010)
2. Crowston, K., Annabi, H., Howison, J.: Defining Open Source Software Project Success. In: 24th International Conference on Information Systems (2003)
3. Crowston, K., Li, Q., Wei, K., Eseryel, U.Y., Howison, J.: Selforganization of teams for free/libre open source software development. Inf. Softw. Technol. 49, 564–575 (2007)
4. Deshpande, A., Riehle, D.: The Total Growth of Open Source. Open Source Development, Communities and Quality, 197–209 (2008)
5. DiBona, C., Ockman, S., Stone, M.: Open Sources: Voices from the Open Source Revolution. O'Reilly Associates, Inc. (1999)
6. Kohlbacher, M.: The Effects of Process Orientation on Customer Satisfaction, Product Quality and Time-Based Performance. In: 29th International Conference of the Strategic Management Society (2009)
7. Mendling, J.: Metrics for Process Models: Empirical Foundations of Verification, Error Prediction, and Guidelines for Correctness. Springer, Heidelberg (2008)
8. Mockus, A., Fielding, R.T., Herbsleb, J.: A case study of open source software development: the Apache server. In: 22nd International Conference on Software Engineering. ACM, Limerick (2000)
9. Raymond, E.S.: The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary. O'Reilly & Associates, Inc. (2001)
10. Reijers, H., van Wijk, S., Mutschler, B., Leurs, M.: BPM in Practice: Who Is Doing What? In: Hull, R., Mendling, J., Tai, S. (eds.) BPM 2010. LNCS, vol. 6336, pp. 45–60. Springer, Heidelberg (2010)
11. Reijers, H.A.: Design and Control of Workflow Processes: Business Process Management for the Service Industry. Springer, Heidelberg (2003)
12. Rubin, V., Günther, C., van der Aalst, W., Kindler, E., van Dongen, B., Schäfer, W.: Process Mining Framework for Software Processes. In: Wang, Q., Pfahl, D., Raffo, D.M. (eds.) ICSP 2007. LNCS, vol. 4470, pp. 169–181. Springer, Heidelberg (2007)
13. Sharma, S., Sugumaran, V., Rajagopalan, B.: A framework for creating hybrid-open source software communities. Information Systems Journal 12, 7–25 (2002)
14. Trung, D.-T., Bieman, J.M.: Open source software development: a case study of FreeBSD. In: 10th International Symposium on Software Metrics, pp. 96–105 (2004)
15. van der Aalst, W.M.P., ter Hofstede, A.H.M., Weske, M.: Business Process Management: A Survey. In: van der Aalst, W.M.P., ter Hofstede, A.H.M., Weske, M. (eds.) BPM 2003. LNCS, vol. 2678, pp. 1–12. Springer, Heidelberg (2003)
16. vom Brocke, J.H., Rosemann, M.: Handbook on Business Process Management: Strategic Alignment, Governance, People and Culture. Springer, Berlin (2010)
17. Zur Muehlen, M.: Workflow-Based Process Controlling. Logos (2004)

Analyzing Stability of Algorithmic Systems Using Algebraic Constructs

Susmit Bagchi

Department of Informatics,
Gyeongsang National University Jinju, South Korea 660 701
susmitbagchi@yahoo.co.uk

Abstract. In general, the modeling and analysis of algorithmic systems involve discrete structural elements. However, the modeling and analysis of recursive algorithmic systems can be done in the form of differential equation following control theoretic approaches. In this paper, the modeling and analysis of generalized algorithmic systems are proposed based on heuristics along with z-domain formulation in order to determine the stability of the systems. The recursive algorithmic systems are analyzed in the form of differential equation for asymptotic analysis. The biplane structure is employed for determining the boundary of the recursions, stability and, oscillatory behaviour. This paper illustrates that biplane structural model can compute the convergence of complex recursive algorithmic systems through periodic perturbation.

Keywords: recursive algorithms, z-domain, stochastic, control theory, perturbation.

1 Introduction

The algorithm design and analysis are the fundamental aspects of any computing systems. The modeling and analysis of algorithms provide an analytical insight along with high-level and precise description of the functionalities of systems [3, 4, 6]. In general, the recursive algorithms are widely employed in many fields including computer-controlled and automated systems [10]. Traditionally, the algorithms are analyzed within the discrete time-domain paying attention to the complexity measures. However, the convergence property and the stability analysis of the algorithms are two important aspects of any algorithmic systems [10]. In case of recursive algorithms, the convergence analysis is often approximated case by case. The asymptotic behaviour of algorithms is difficult to formulate with generalization [4, 10]. The asymptotic behaviour of stochastic recursive algorithms is formulated by constructing models [10], however, such models fail to analyze the stability of the algorithm in continuous time domain throughout the execution. This paper argues that the stability analysis of any algorithm can be performed within the frequency-domain by considering the algorithms as functional building blocks having different configurations. In order to perform generalized frequency-domain analysis, the algorithms are required to be modeled and transformed following the algebraic

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 81–90, 2013.

constructs. Furthermore, this paper proposes that boundary of execution of recursive algorithms can be analyzed following biplane structure and the stability of the algorithms can be observed in the presence of stochastic input by following the traces in the biplane structure bounding the algorithms. The proposed analytical models are generalized without any specific assumptions about the systems and thus, are applicable to wide array of algorithmic systems. This paper illustrates the mechanism to construct analytical model of any complex algorithmic system and methods to analyze the stability of the system under consideration. The rest of the paper is organized as follows. Section 2 describes related work. Section 3 illustrates the modeling and analysis of the algorithms in frequency-domains and their stability analysis using biplane structure. Section 4 and 5 present discussion and conclusion, respectively.

2 Related Work

The modeling of computer systems and algorithms is useful to gain an insight to the designs as well as to analyze the inherent properties of the systems [2, 3, 4, 6, 7, 10]. For example, the fusion of models of artificial neural network (ANN) and fuzzy inference systems (FIS) are employed in many complex computing systems. The individual models of the ANN and FIS are constructed and their interactions are analyzed in order to establish a set of advantages and disadvantages overcoming the complexities of these systems [1]. The other successful applications of modeling techniques to the distributed algorithms and the distributed database in view of Petri Nets are represented in [2, 6]. It is illustrated how Petri Nets can be employed to model and analyze complex distributed computing algorithms [2]. However, in case of distributed database, the concurrency control algorithms are modeled by formulating extended place/transition net (EPTN) [6]. The EPTN formalism is a derivative of the Petri Nets. In structured peer-to-peer (P2P) networks, the random-walks mechanism is used to implement searching of information in minimum time. The model of searching by random-walks in P2P network is constructed to obtain analytical expressions representing performance metrics [3]. Following the model, an equation-based adaptive search in P2P network is presented. The analysis of probabilistic as well as real-time behaviour and the correctness of execution are the challenges of systems involving wireless sensor networks (WSN). Researchers have proposed the modeling techniques of WSN to analyze the behaviour, correctness and performance of WSN by using Real-Time Maude [4]. The Real-Time Maude model provides an expressive tool to perform reachability analysis and the checking of temporal logic in WSN systems. On the other hand, the modeling and analysis of hand-off algorithms for cellular communication network are constructed by employing various modeling formalisms [5, 8]. The modeling of fast hand-off algorithms for microcellular network is derived by using the local averaging method [5]. The performance metrics of the fast hand-off algorithms and the necessary conditions of cellular structures are formulated by using the model construction. In another approach, the modeling technique is employed to evaluate the hand-off algorithms for cellular network [8]. In this case, the model is constructed based on the

estimation of Wrong Decision Probability (WDP) and the hand-off probability [8]. In the image processing systems, the modeling and analysis of signals are performed by designing the sliding window algorithms. Researchers have proposed the Windowed Synchronous Data Flow (WSDF) model to analyze the sliding window algorithms [7]. The WSDF is constructed as a static model and a WSDF-balance equation is derived.

The analysis of convergence of any algorithm is an important phenomenon [9, 10]. The convergence analysis of canonical genetic algorithms is analyzed by using modeling techniques based on homogeneous finite Markov chain [9]. The constructed model illustrates the impossibility of the convergence of canonical genetic algorithms towards global optima. The model is discussed with respect to the schema theorem. On the other hand, the modeling and analysis of generalized stochastic recursive algorithms are performed using heuristics [10]. The heuristic model explains the asymptotic behaviour of stochastic recursive algorithms. However, the model does not perform the stability analysis of the recursive algorithms in the presence of stochastic input.

3 Models of Algorithms in z-domain

The z-domain analysis is widely used to analyze the dynamics and stability of the discrete systems. The computing algorithms can be modeled in z-domain in order to construct heuristic analysis as well as stability analysis of the various algorithmic models in the view of the transfer functions.

3.1 Singular Model

In the singular model, the algorithm is considered as a transfer function with single input and single output (SISO) mechanism. The schematic representation of the singular model is presented in Fig. 1.

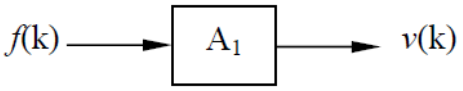


Fig. 1. Schematic representation of singular model

In SISO model, the algorithm A_1 acts as a discrete transfer function for instances k = 0, 1, 2, 3, …..N and transfers the discrete input $f(k)$ into corresponding discrete output $v(k)$. Let, a non-commutative composition of any two functions x and y is described as $(x_o y)$. Thus, the dynamics of the singular algorithmic model can be composed as, $v(k) = A_1(f(k)) = (A_{1o}f)(k)$. Let, $\alpha_1 = (A_{1o}f)$, hence in z-domain $v(z) = \Sigma_{k=0,\infty}\, \alpha_1(k).z^{-k} = \alpha_1(z)$. The algorithmic transfer function is stable if $\alpha_1(z)$ is a monotonically decreasing function for sufficiently large k.

3.2 Chained Model

In the chained model of the algorithmic system, two independent algorithms are put in series as illustrated in Fig. 2.

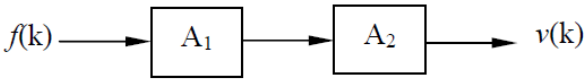


Fig. 2. Schematic representation of chained model

In the chained model, two algorithms act as independent transfer functions transforming discrete input to discrete output at every instant k. Thus, the overall transfer function of chained model can be presented as, $v(k) = (A_{2o}\alpha_1)(k) = \alpha_{21}(k)$. Hence, in the z-domain $v(z) = \alpha_{21}(z)$ and the chained algorithms are stable if $\alpha_{21}(z)$ is monotonically decreasing for sufficiently large k.

3.3 Parallel Models

In case of parallel model, two (or more) independent algorithms execute in parallel on a single set of input at every instant k and, the final output of the system is composed by combining the individual outputs of the algorithms. A 2-algorithms parallel model is illustrated in Fig. 3.

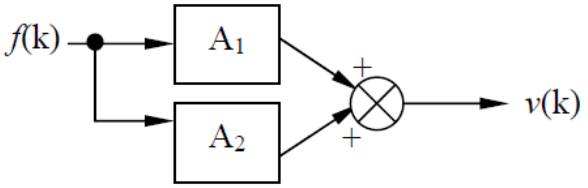


Fig. 3. Schematic representation of 2-algorithms parallel model

So, in the 2-algorithms parallel model, the output is computed as, $v(k) = (A_{1o}f)(k) + (A_{2o}f)(k) = \alpha_1(k) + \alpha_2(k)$. Hence, in the z-domain the discrete values of the output can be presented as, $v(z) = \alpha_1(z) + \alpha_2(z)$. This indicates that a parallel algorithmic system is stable if either the individual algorithmic transfer functions are monotonically decreasing or the combined transfer function of the system is converging for sufficiently large k. On the other hand, the 2-algorithms parallel model can be further extended to parallel-series model by adding another algorithm in series as illustrated in Fig. 4.

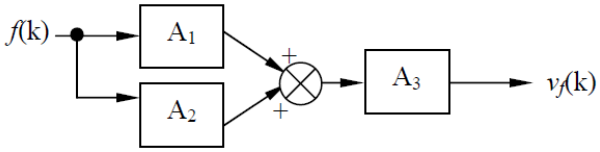


Fig. 4. Schematic representation of 2-algorithms parallel-series model

However, in this case algorithm A_3 transforms the output of parallel computation in deterministic execution at discrete instances k. Hence, the final output of the system is, $v_f(k) = A_3(\alpha_1(k) + \alpha_2(k))$. As, $v(k) = \alpha_1(k) + \alpha_2(k)$, thus $v_f(k) = \alpha_3(k)$, where $\alpha_3 = (A_3 \circ v)$ and $v_f(z) = \alpha_3(z)$. The parallel-series model is stable if $\alpha_3(z)$ is a converging function. This indicates that, $v_f(z)$ can be stable even if $v(z)$ is diverging function provided $A_3(v(z))$ is a monotonically converging function.

3.4 Recursion with Stochastic Observation

The recursive algorithms are widely used in computing systems. The fundamental aspect of any computing system involving the recursive algorithm is the existence of a feedback path as illustrated in Fig. 5. In the feedback algorithmic model, the feedback path is positive and the feedback gain can be either unity or can have any arbitrary transfer-gain.

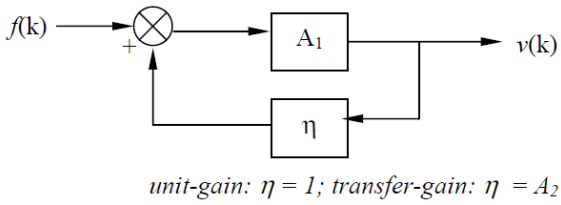


Fig. 5. Schematic representation of recursive model

In pure recursive computing model, the feedback path will have unit gain and the system will be controlled by external input $f(k)$ at $k = 0$, whereas the further input series to the system will change to $v(k)$ due to positive feedback for $k > 0$, where $f(k > 0) = 0$. The behaviour of such system can be analyzed by using two formalisms such as, heuristics and z-domain analysis techniques.

3.4.1 Heuristics Analysis

The generalized difference equation of the recursive algorithmic system is given as, $v(k) = A_1(A_2(v(k\text{-}1)) + f(k))$. In case of positive feedback with unit gain, the closed-loop difference equation controlling the system is given as,

$$v(k) = A_1(v(k\text{-}1) + f(k)) \tag{1}$$

Equation (1) represents a recursive algorithm with stochastic input $f(k)$, where A_1 is a deterministic function having a set of regularity conditions. The function $f(k)$ can be generalized by using some function y as,

$$f(k) = y(v(k-1), f(k-1), e(k)) \tag{2}$$

where, $e(k)$ is an arbitrary error in the system at the execution instant k.

The stability of whole system depends on the stability of equation (2). If $f(k)$ is exponentially stable within a small neighborhood around k after some point b ($k >> b$), then [10],

$$B(v(b-1), f(b)) = h(v(k)) + r(b) \tag{3}$$

where, $B(v(k-1), f(k)) = A_1(v(k-1) + f(k))$, $h(v(\ )) = EB(v, \overline{f}(b))$ and $r(b)$ is a random variable with zero mean. Thus, equation (1) can be represented as,

$$v(k) = B(v(k-1), f(k)) \tag{4}$$

Hence, equation (4) can be approximately evaluated between k and k+a ($a > 0$) as,

$$\begin{aligned} v(k+a) &= v(k) + \Sigma_{j = k+1,\, k+a}\, B(v(j-1), f(j)) \\ &\approx v(k) + \Sigma_{j = k+1,\, k+a}\, h(v(k)) + \Sigma_{j = k+1,\, k+a}\, r(j) \\ &\approx v(k) + a\, h(v(k)) \end{aligned} \tag{5}$$

In equation (5), the random variable is eliminated as it has the zero mean. Hence, the differential equation at point a is given by,

$$lim_{a \to 0}[v(k+a) - v(k)]/a = dv(k)/da = h(v(k)) \tag{6}$$

Thus, the asymptotic properties of the equation (1) can be further derived from equation (6) in the form of derivative for any specific recursive algorithmic system.

3.4.2 Stability in z-domain

For the stability analysis in z-domain, it is assumed that $A_1(k)$ represents the gain factor of A_1 at k-th instant of execution of the algorithm. Now, $v(k) = (\eta(v(k-1)) + f(k))A_1(k)$, where $f(k)$ is a singular external input to A_1 defined as, $f(k) = m$ if $k = 0$ and, $f(k) = 0$ otherwise. Hence, $v(k) = \eta A_1(k)v(k-1) + f(k)A_1(k)$. Initially at $k = 0$, $v(0) = mA_1(0)$. Hence, $v(k) = \eta A_1(k)v(k-1) + mA_1(0)$. If the system is purely recursive, then feedback gain is unity ($\eta = 1$) and, $v(k) = A_1(k)v(k-1) + mA_1(0)$. Thus, in the z-domain the system will be represented as, $v(z) = mA_1(0)z/(z-1) + \Sigma_{k = 2, \infty}\, A_1(k)\, v(k-1)z^{-k}$. Deriving further one can get,

$$\begin{aligned} v(z) &= mA_1(0)z/(z-1) + \{A_1(1)v(0)/z + A_1(2)\, v(1)/z^2 + \ldots\ldots..\} \\ &= mA_1(0)z/(z-1) + mA_1(0)\, \Sigma_{k=1,\infty}\, A_1(k)\, z^{-k} + \\ &\quad \Sigma_{k=2,\infty}\, \{\Pi_{j=k,k-1}\, A_1(j)\}v(k-2)\, z^{-k} \\ &= mA_1(0)z/(z-1) + mA_1(0)[A_1(z) - A_1(0)] + \Lambda_z \end{aligned} \quad (7)$$

where, $\Lambda_z = \Sigma_{k=2,\infty}\, \{\Pi_{j=k,k-1}\, A_1(j)\}v(k-2)\, z^{-k}$.

The system will be stable if Λ_z will minimize or converge for sufficiently large k.

3.5 Functional Properties

The functional properties of a generalized recursive algorithm with unit positive feedback analyze the stability of the overall system in the presence of oscillation, if any. In addition, the concept of biplane symmetry can be used to analyze the bounds of a recursive algorithmic system. The generalized recursive algorithmic model with positive transfer-gain is represented as $v(k) = A_1(A_2(v(k-1)) + f(k))$. Let, $(A_{1o}A_2) = \delta$ and $f(k)$ is a singular external input to algorithm defined as, $f(k) = m$ if $k = 0$ and, $f(k) = 0$ otherwise. Thus, the initial output value is $v(0) = A_1(m)$ and $v(k) = \delta^k(d)$, where $d = A_1(m)$. Now, if A_2 is a unit gain factor, then the system reduces to a pure recursive algorithm such that, $v(k) = A_1^{\,k}(d)$, $k > 0$. The stability and behavioral properties of the recursive algorithmic system can be further analyzed as follows.

3.5.1 Stability and Convergence

Let, $\not{f}: \mathcal{R} \rightarrow \mathcal{R}$ is a stochastic function defined on space $\mathcal{R}$ such that, $\delta(d) \in \not{f}(\mathcal{R}) \subset \mathcal{R}$ and $|\not{f}(\mathcal{R})| > 1$. Now, for $k > 0$, the $\delta^k(d) \in \not{f}^k(\mathcal{R})$ such that, either $\not{f}^k(\mathcal{R}) \cap \not{f}^{k+1}(\mathcal{R}) = \{\phi\}$ or $\not{f}^k(\mathcal{R}) \cap \not{f}^{k+1}(\mathcal{R}) \neq \{\phi\}$ depending on the dynamics. A system is bounded if $\not{f}^{k+1}(\mathcal{R}) \subseteq \not{f}^k(\mathcal{R})$. The boundary of $\delta^k(d)$ is $\Delta_k = \cap_{i=1,k}\, \not{f}^i(\mathcal{R})$. A ε-cut of $\not{f}^k(\mathcal{R})$ is defined as $\not{f}_{k\varepsilon} \subset \not{f}^k(\mathcal{R})$ such that, $\forall a \in \not{f}_{k\varepsilon}$ the following condition is satisfied: $\varepsilon \in \not{f}^k(\mathcal{R})$ and $a > \varepsilon$. An instantaneous remainder of $\not{f}^k(\mathcal{R})$ is given by, $\overline{\not{f}}_{k\varepsilon} = (\not{f}^k(\mathcal{R}) - \not{f}_{k\varepsilon})$. A system is stable at point N if the boundary $\Delta_N \neq \{\phi\}$, where $1 \leq |\Delta_N| \leq w$ and $w << N$. A converging system is a stable system at recursion level N with $|\Delta_N| = 1$.

3.5.2 Divergence in Systems

Let, in a system $\delta^{k-1}(d) \in \not{f}_{(k-1)\varepsilon}$ whereas $\delta^k(d) \in \not{f}_{k\varepsilon}$ and $\delta^{k+1}(d) \in \not{f}_{(k+1)\varepsilon}$ such that, $\delta^{k-1}(d) < \delta^k(d) < \delta^{k+1}(d)$. The system is divergent if $\not{f}_{(k-1)\varepsilon} \cap \not{f}_{k\varepsilon} \cap \not{f}_{(k+1)\varepsilon} = \{\phi\}$. A divergent system is unstable if the limit of recursion $k >> 1$.

3.5.3 Biplane Symmetries

Let, in a system for $k \geq 1$, $\not{f}^k(\mathcal{R}) = \not{f}(\mathcal{R})$ and, $\not{f}^*: \mathcal{R} \rightarrow \mathcal{R}$ such that $(\not{f}^*)^k(\mathcal{R}) = \not{f}^*(\mathcal{R})$ where, $\not{f}(\mathcal{R}) \cap \not{f}^*(\mathcal{R}) = \{\phi\}$. Furthermore, $\not{f}_{*\varepsilon}$ is the ε-cut of $\not{f}^*(\mathcal{R})$ and $\not{f}_\varepsilon$ is the ε-cut of $\not{f}(\mathcal{R})$, whereas the corresponding remainders are $\overline{\not{f}}_{*\varepsilon}$ and $\overline{\not{f}}_\varepsilon$, respectively. Let, $\delta^p(d) \in \not{f}(\mathcal{R})$ for $p = 1, 3, 5, \ldots..$ and, $\delta^q(d) \in \not{f}^*(\mathcal{R})$ for $q = p + 1$. Now, if $x_{p+j} = \delta^{p+j}(d)$ and $y_{q+j} = \delta^{q+j}(d)$, $j = 0, 2, 4, 6\ldots\ldots.$, then following set of predicates can occur in the system,

$P1 \Rightarrow [(x_p \in \overline{\not{f}_\varepsilon}) \wedge (x_{p+2} \in \not{f}_{*\varepsilon}) \wedge (x_{p+4} \in \overline{\not{f}_\varepsilon}) \wedge \ldots\ldots\ldots..]$
$P2 \Rightarrow [(x_p \in \overline{\not{f}_\varepsilon}) \wedge (x_{p+2} \in \not{f}_\varepsilon) \wedge (x_{p+4} \in \overline{\not{f}_\varepsilon}) \wedge \ldots\ldots\ldots\ldots]$
$P3 \Rightarrow [(y_q \in \overline{\not{f}_{*\varepsilon}}) \wedge (y_{q+2} \in \not{f}_{*\varepsilon}) \wedge (y_{q+4} \in \overline{\not{f}_{*\varepsilon}}) \wedge \ldots\ldots\ldots]$
$P4 \Rightarrow (x_{p+j} \in \overline{\not{f}_\varepsilon})$
$P5 \Rightarrow (x_{p+j} \in \not{f}_\varepsilon)$
$P6 \Rightarrow (y_{q+j} \in \overline{\not{f}_{*\varepsilon}})$
$P7 \Rightarrow (y_{q+j} \in \not{f}_{*\varepsilon})$

The possible combinatorial distributions of predicates in a recursive algorithmic system are, P_{13}, P_{23}, P_{46}, P_{47}, P_{56}, P_{57} where, $P_{ab} = (Pa \wedge Pb)$. If distributions P_{46} and P_{57} are valid in a recursive algorithmic system, then it is a biplane-symmetric algorithmic system. Otherwise, if the distributions P_{47} and P_{56} are valid in a system, then the system is a biplane-asymmetric system. Furthermore, if the distribution P_{23} is satisfied in a recursive algorithmic system, then the system is having dual-symmetry between biplanes $\not{f}$ and $\not{f}^*$ and the system is represented as, $[\not{f}/\not{f}^*]$. On the other hand, if the distribution P_{13} is satisfied in a recursive algorithmic system, then the system is called Bounded-Periodic-Perturbed (BPP) system represented as $(\not{f}_{*\varepsilon}|$. In this case, the system is bounded within $\not{f}$ and $\not{f}^*$ planes, however periodic perturbations occur within the domain $\not{f}_{*\varepsilon}$.

3.5.4 Oscillation in Recursive Systems

In a biplane-symmetric system if the following properties hold, then it is called the biplane-symmetric oscillatory recursive system, $\forall p, q, |x_p| = |y_q| = |x_{p+j}| = |y_{q+j}|$ and $x_p + y_q = x_{p+j} + y_{q+j} = 0$. However, in a $[\not{f}/\not{f}^*]$ system if the following conditions hold, then the system is called asymmetrically oscillating between $\not{f}$ and $\not{f}^*$ planes for values of s (s = 0, 4, 6……), $\forall p, q, |x_{p+s}| = |y_{q+s}|, |x_{p+s+2}| = |y_{q+s+2}|$ and $x_{p+s} + y_{q+s} = 0$, $x_{p+s+2} + y_{q+s+2} = 0$. If a recursive algorithmic system is oscillatory, then it is a deterministic but non-converging system.

A recursive algorithmic system is deterministic and converging if there exists a constant *C* such that, $\sum_{p = 1, N} (x_p + y_{p+1}) = \sum_{p = 1, M} (x_p + y_{p+1}) = C$, where $N \neq M$. This indicates that, a deterministic and converging recursive algorithmic system should be in damped oscillation (stable) and should contain idempotency. On the other hand, an oscillatory non-converging recursive algorithmic system is non-idempotent requiring strict consistency conditions.

4 Discussion

Traditionally, the recursive algorithmic systems are analyzed by using heuristics as well as asymptotic methods following difference equation. Often, the differential equation is formulated in place of difference equation in order to conduct analysis in continuous plane avoiding the complexity. However, the generalized z-domain analysis of algorithmic systems in a discrete plane offers an insight towards the understanding of the overall stability of the algorithmic systems.

The perturbation analysis of a system using biplane structure captures the inherent oscillation in the system. The determinism of convergence of the recursive algorithmic systems with stochastic input can be computed using the symmetry of biplane structure. As a result, the idempotent property of the recursive algorithmic systems becomes easily verifiable in case of complex systems. Thus, depending upon the idempotent property of the complex recursive algorithmic systems, the appropriate consistency conditions can be designed.

5 Conclusion

The analysis of stability and behaviour of any algorithmic systems can be accomplished by modeling such systems as a block having transfer functional properties. The z-domain analysis of algorithmic models captures the overall response trajectory and the stability of the algorithmic systems. The complex recursive algorithmic systems can be analyzed by modeling in view of z-domain and biplane structure. The heuristics and z-domain models of a generalized recursive algorithmic system with stochastic input reduce the overall system to the deferential equation presenting the dynamic behaviour of the recursive algorithm. On the other hand, the biplane structure determines the boundaries of the recursive algorithmic systems. In addition, the biplane structural model of recursive algorithmic systems serves as a tool to analyze the oscillatory nature of the recursions as well as the stability of the algorithmic systems. The biplane structural model helps to achieve periodic perturbation into the system dynamics and determining convergence conditions, which enables to design the appropriate consistency conditions.

References

1. Abraham, A.: Neuro Fuzzy Systems: Sate-of-the-Art Modeling Techniques. In: Mira, J., Prieto, A.G. (eds.) IWANN 2001. LNCS, vol. 2084, pp. 269–276. Springer, Heidelberg (2001)
2. Reisig, W.: Elements of distributed algorithms: modeling and analysis with petri nets. Springer (1998)
3. Bisnik, N., Abouzeid, A.: Modeling and analysis of random walk search algorithms in P2P networks. In: Second International Workshop on Hot Topics in Peer-to-Peer Systems (HOT-P2P 2005), pp. 95–103. IEEE (2005)
4. Olveczky, P.C., Thorvaldsen, S.: Formal modeling and analysis of wireless sensor network algorithms in Real-Time Maude. In: 20th International Symposium on Parallel and Distributed Processing (IPDPS 2006). IEEE (2006)
5. Leu, A.E., Mark, B.L.: Modeling and analysis of fast handoff algorithms for microcellular networks. In: 10th IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunications Systems (MASCOTS 2002). IEEE (2002)
6. Ozsu, M.T.: Modeling and Analysis of Distributed Database Concurrency Control Algorithms Using an Extended Petri Net Formalism. IEEE Transactions on Software Engineering SE-11(10) (1985)

7. Keinert, J., Haubelt, C., Teich, J.: Modeling and Analysis of Windowed Synchronous Algorithms. In: 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006). IEEE (2006)
8. Chi, C., Cai, X., Hao, R., Liu, F.: Modeling and Analysis of Handover Algorithms, Global Telecommunications Conference (GLOBECOM 2007). IEEE (2007)
9. Rudolph, G.: Convergence analysis of canonical genetic algorithms. IEEE Transactions on Neural Networks 5(1) (1994)
10. Ljung, L.: Analysis of recursive stochastic algorithms. IEEE Transactions on Automatic Control 22(4) (1977)

Algorithms of the Combination of Compiler Optimization Options for Automatic Performance Tuning

Suprapto and Retantyo Wardoyo

Department of Computer Science and Electronics
Faculty of Mathematics and Natural Sciences,
Universitas Gadjah Mada Yogyakarta, Indonesia
{sprapto,rw}@ugm.ac.id

Abstract. It is very natural when people compile their programs, they would require a compiler that gives the best program performance. Even though today's compiler have reached the point in which they provide the users a large number of options, however, because of the unavailability of program input data and insufficient knowledge of the target architecture; it can still seriously limit the accuracy of compile-time performance models. Thus, the problem is how to choose the best combination of optimization options provided by compiler for a given program or program section. This gives rise the requirement of an orchestration algorithm that fast and effective to search for the best optimization combination for a program.

There have been several algorithms developed, such as Exhaustive Search (ES); Batch Elimination (BE); Iterative Elimination (IE); Combined Elimination (CE); Optimization Space Exploration (OSE); and Statistical Selection (SS). Based on those of algorithms, in this paper we proposed Heuristics Elimination (HE) algorithm, a simple algorithm that was mostly inspired by OSE with some differences. The HE algorithm uses a heuristic approach by applying genetic algorithm to find the best combination of compiler's optimization options. It is unlike OSE, however, this proposed algorithm starts from a set of some possible combinations randomly selected, then they are iteratively refined by some genetic operators to find one optimal combination (as the solution).

Keywords: Compiler optimization, optimization options, performance, orchestration algorithm, exhaustive search, batch elimination, iterative elimination, combined elimination, optimization space exploration, statistical selection.

1 Introduction

As we all know that optimizations of compiler for modern architectures have achieved high level of sophistication [11]. Although compiler optimizations have made a significant improvements in many programs, however, the potential for the degradation of performance in certain program patterns is still seen by compiler developer and many users. The state of the art is allowing the users deal

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 91–100, 2013.

with this problem by providing them many compiler options. This compiler options's existence indicates that today's optimizers are not capable of making optimal choices at compile time. Moreover, the availability of input data of program is very minimum, and the lack of knowledge about the target architecture can limit the accuracy of compile-time performance models.

Therefore, the determination of the best combination of compiler optimizations for a given program or program section remains an unattainable compile-time goal. Today's compilers have evolved to the situation in which users are provided with a large number of options. For instance, GCC Compilers include 38 options, roughly grouped into three optimization levels, O1 through O3 [11]. On the other hand, compiler optimizations interact in unpredictable manners, as many have observed [2], [4], [10], [11], [8], [9]. Therefore, it is desired a fast and effective orchestration algorithm to search for the best optimization combination for a program.

Many automatic performance tuning systems have taken a dynamic, feedback-directed approach to orchestra compiler optimizations. In this approach, many different binary code versions generated under different experimental optimization combinations are being evaluated. The performance of these versions is compared using either measured execution times or profile-based estimates. Iteratively, the orchestration algorithms use this information to decide the next experimental optimization combinations, until converge criteria are reached [11].

2 Algorithms of Orchestration

In this section, we briefly present an overview of some algorithms that have goal finding an optimal combination of compiler's options. To do this, let we first define the goal of optimization orchestration as follows :

Given a set of compiler optimization options $\{F_1, F_2, \ldots, F_n\}$, where n is the number of optimization. Find the combination that minimizes the execution time of program efficiently, without using a priori knowledge of the optimization and their interactions.

2.1 Algorithm of Exhaustive Search

The exhaustive search (ES) approach, which is called the *factorial design* [2], [8], would try to evaluate every possible compiler's options in finding the best. This approach provides an upper bound of an application's performance after optimization orchestration. However, its complexity $O(2^n)$, which is prohibitive if it involves a large number of compiler's options. As an illustration, for GCC compiler with 38 options, it would take up to 2^{38} program runs – a million years is required for a program that runs in two minutes. Considering this fact, this algorithm will not be evaluated under the full set of options [11]. By the use of pseudo code, ES can be depicted as follows.

1. Get all 2^n combination of n compiler's options, $\{F_1, F_2, \ldots, F_n\}$.
2. For the optimized version compiled under every combination of n compiler's options, measure application execution time.

3. An optimal combination of compiler's options is one that give the smallest execution time to the program.

2.2 Algorithm of Batch Elimination

The idea of Batch Elimination (BE) is to identify the optimizations with negative effects and turn them off at once. BE achieves good program performance, when the compiler's options do not interact each other [11], [12], and [13]. The negative effect of one compiler's option, F_i can be represented by its *RIP* (*Relative Improvement Percentage*), $RIP(F_i)$, (see equation 1) which is the relative difference of the execution times of the two versions with and without F_i, that means $T(F_i = 1)$ and $T(F_i = 0)$ respectively ($F_i = 1$ means F_i is on, and $F_i = 0$ means F_i is off).

$$RIP(F_i) = \frac{T(F_i = 0) - T(F_i = 1)}{T(F_i = 1)} \times 100\% \tag{1}$$

The baseline of this approach switches on all compiler optimization options. $T(F_i = 1)$ is the execution time of the baseline T_B as shown in equation 2.

$$T_B = T(F_i = 1) = T(F_1 = 1, \ldots, F_n = 1) \tag{2}$$

The performance improvement by switching off F_i from the baseline B relative to the baseline performance can be calculated with equation 3.

$$RIP_B(F_i = 0) = \frac{T(F_i = 0) - T_B}{T_B} \times 100\% \tag{3}$$

If $RIP_B(F_i = 0) < 0$, the optimization of F_i has a negative effect. The BE algorithm eliminates the optimizations with negatives RIP in a batch to generate the final combination tuned version. The complexity of BE algorithm is $O(n)$.

1. Compile the application under the baseline $B = \{F_1, \ldots, F_n\}$. Execute the generated code version to obtain the baseline execution time T_B.
2. For each optimization F_i, switch it off from B and compile the application. Execute the generated version to obtain $T(F_i = 0)$, and compute $RIP_B(F_i = 0)$ according to equation 3.
3. Disable all optimizations with negative RIP to generate the final tuned version.

2.3 Algorithm of Iterative Elimination

Iterative Elimination (IE) algorithm was designed to consider the interaction of optimizations. Unlike BE algorithm, which turns off all optimizations with negative effects at once, IE algorithm iteratively turns off one optimization with the most negative effect at a time.

IE algorithm starts with the baseline that switches all compiler's optimization options **on** [11], and [7]. After computing the RIPs of the optimizations according to equation 3, IE switches the one optimization with the most negative effect **off** from the baseline. This process repeats with all remaining optimizations, until none of them causes performance degradation.

1. Let B be the combination of compiler optimization options for measuring the baseline execution time, T_B. Let the set S represent the optimization searching space. Initialize $S = \{F_1, \ldots, F_n\}$ and $B = \{F_1 = 1, \ldots, F_n = 1\}$.
2. Compile and execute the application under the baseline setting to obtain the baseline execution time T_B.
3. For each optimization option $F_i \in S$, switch F_i **off** from the baseline B and compile the application, execute the generated code version to obtain $T(F_i = 0)$, and compute the RIP of F_i relative to the baseline B, $RIP_B(F_i = 0)$, according to equation 3.
4. Find the optimization F_x with the most negative RIP. Remove F_x from S, and set F_x to 0 in B.
5. Repeat Steps 2, 3, and 4 until all options in S have non-negative RIPs. B represent the final option combination.

2.4 Algorithm of Combined Elimination

Combined Elimination (CE) algorithm combines the ideas of the two algorithms (BE and IE) just described [11], and [7]. It has a similar iterative structure as IE. In each iteration, however, CE applies the idea of BE : after identifying the optimization with negative effects (in this iteration), CE tries to eliminate these optimizations one by one in a greedy fashion.

Since IE considers the interaction of optimizations, it achieves better performance of program than BE does. When the interactions have only small effects, however, BE may perform close to IE in a faster way. Based on the way CE designed, it takes the advantages of both BE and IE. When the optimizations interact weakly, CE eliminates the optimizations with negative effects in one iteration, just like BE. Otherwise, CE eliminates them iteratively, like IE. As a result, CE achieves both good program performance and fast tuning speed.

1. Let B be the baseline option combination, and let the set S represent the optimization search space. Initialize $S = \{F_1, \ldots, F_n\}$ and $B = \{F_1 = 1, \ldots, F_n = 1\}$.
2. Compile and execute the application under baseline setting to obtain the baseline execution time T_B. Measure the $RIP_B(F_i = 0)$ of each optimization options F_i in S relative to the baseline B.
3. Let $X = \{X_1, \ldots, X_l\}$ be the set of optimization options F_i with negative RIPs. X is stored in increasing order, that is, the first element, X_1, has the most negative RIP. Remove X_1 from S, and set X_1 to 0 in the baseline B (in this step, B is updated by setting the optimization option with the most negative RIP to zero). For i from 2 to l,

- Measure the RIP of X_i relative to the baseline B.
- If the RIP of X_i is negative, remove X_i from S and set X_i to 0 in B.

4. Repeat Steps 2 and 3 until all options in S have non-negative RIPs. B represents the final solution.

2.5 Algorithm of Optimization Space Exploration

The basic idea of algorithm ***pruning*** is to iteratively find better combination of optimization options by merging the beneficial ones [9]. In each iteration, a new test set Ω is constructed by merging the combination of optimization options in the old test set using **union** operation. Then, after evaluating the combination of optimization options in Ω, the size of Ω is reduced to m by dropping the slowest combinations. The process repeats until the performance increase in the Ω set of two consecutive iteration become negligible. The specific steps are as follows :

1. Construct a set, Ω, which consists of the default optimization combination, and n combinations, each of which assigns a non-default value to a single optimization. In the experiment [11], the default optimization combination, O3,turns **on** all optimizations. The non-default value for each optimization is **off**.
2. Measure the application execution time for each optimization combination in Ω. Keep the m fastest combination in Ω, and remove the rest (i.e., n - m combinations).
3. Construct a new set of Ω, each element in which is a union of two optimization combinations in the old Ω. (The **union** operation takes non-default values of the options in both combinations.)
4. Repeats Steps 2 and 3, until no new combinations can be generated or the increase of the fastest version in Ω becomes negligible. The fastest version in the final Ω as the final version.

2.6 Algorithm of Statistical Selection

Statistical Selection (SS) algorithm uses a statistical method to identify the performance effect of the optimization options. The options with positive effects are turned **on**, while the one with negative effects are turned **off** in the final version, in an iterative fashion. This statistical method takes the interaction of optimization options into consideration.

The statistical method is based on orthogonal arrays (OA), which have been proposed as an efficient design of experiments [3], [1]. Formally, an OA is an $m \times k$ matrix of zeros and ones. Each column of the array corresponds to one compiler option, while each row of the array corresponds to one optimization combination. SS algorithm uses the OA with strength 2, that is, two arbitrary columns of the OA contain the patterns 00, 01, 10, 11 equally often. The experiments [11] used the OA with 38 options and 40 rows, which is constructed based on a *Hadamard* matrix taken from [6].

By a series of program runs, this SS algorithm identifies the options that have the largest effect on code performance. Then, it switches on/off those options with a large positive/negative effect. After iteratively applying the above solution to the options that have not been set, SS algorithm finds an optimal combination of optimization options. The pseudo code is as follows.

1. Compile the application with each row from orthogonal array A as the combination of compiler optimization options, and execute of the optimized version.
2. Compute the *relative effect*, $RE(F_i)$, of each option using equation 4 and 5, where $E(F_i)$ is the *main effect* of F_i, s is one row of A, $T(s)$ is the execution time of the version under s.

$$E(F_i) = \frac{(\sum_{s \in A:s_i=1} T(s) - \sum_{s \in A:s_i=0} T(s))^2}{m} \tag{4}$$

$$RE(F_i) = \frac{E(F_i)}{\sum_{j=1}^{k} E(F_j)} \times 100\% \tag{5}$$

3. If the relative effect of an option is greater than a threshold of 10%.
   - if the option has a positive *improvement*, $I(F_i) > 0$, according to equation 6, switch the option **on**,
   - else if the option has a negative *improvement*, switch the option **off**.

$$I(F_i) = \frac{\sum_{s \in A:s_i=0} T(s) - \sum_{s \in A:s_i=1} T(s)}{\sum_{s \in A:s_i=0} T(s)} \tag{6}$$

4. Construct a new orthogonal array A by dropping the columns corresponding to the options selected in the previous step.
5. Repeat all above steps until all of the options are set.

3 Algorithm of Heuristics Elimination

As mentioned in the previous section, *Heuristic Elimination* (HE) algorithm was mostly inspired by OSE algorithm; that is why the way it works is similar to OSE algorithm with some differences. HE algorithm iteratively find the combination of compiler's optimization options by applying heuristic approach using genetic algorithm.

The basic genetic algorithm is very generic and there many aspects that can be implemented very differently according to the problem [5]. For instance, representation of solution or chromosomes, type of encoding, selection strategy, type of crossover and mutation operators, etc. In practice, genetic algorithms are implemented by having arrays of bits or characters to represent the chromosomes. How to encode a solution of the problem into a chromosome is a key issue when using genetic algorithms. The individuals in the populations then go through a process of simulated evolution. Simple bit manipulation operations allow the

implementation of crossover, mutation and other operations. Individual for producing offspring are chosen using a selection strategy after evaluating the fitness value of each individual in the selection pool. Each individual in the selection pool receives a reproduction probability depending on its own fitness value and the fitness value of all other individuals in the selection pool. This fitness is used for the actual selection step afterwards. Some of the popular selection schemes are Roulette Wheel, Tournament, etc.

Crossover and **mutation** are two basic operators of genetic algorithm, and the performance of genetic algorithm very much depends on these genetic operators. Type and implementation of operators depends on encoding and also on the problem [5]. A new population is formed by selecting the fitter individual from the parent population and the offspring population (elitism). After several generations (iterations), the algorithm converges to the best individual, which hopefully represents an optimal or suboptimal solution to the problem.

Given the set of optimization options $\{F_1, F_2, \ldots, F_n\}$, there are exist 2^n possible combinations. It is unlike ES algorithm that evaluate every possible combination of optimization options, however, HE algorithm only needs to evaluate a certain number of combinations, and known as population size (m) of each generation, for instance $\{C_1, C_2, \ldots, C_m\}$. The optimal combination of optimization options is obtained by improving their fitness value iteratively with genetic operators in each generation (or iteration), and the combination with optimal fitness in the last generation will be final (an optimal) solution.

Chromosome that represents the solution of the problem (i.e., the combination of compiler's optimization options) is defined as one has fixed length n, and the value in each location indicates the participation (or involvement) of the compiler's option; which is formally defined in equation 7.

$$F_i = \begin{cases} 1 & \text{if } F_i \text{ involved} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where, $F_i = 1$ for some i if the *i-th* compiler optimization option F_i is involved in the combination; otherwise $F_i = 0$.

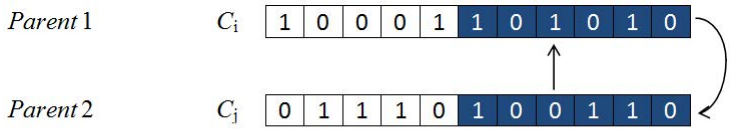
According to equation 7, then chromosome is encoded as a sequence of 0's or 1's (binary representation). For instance, 01100011110101 ($n = 14$) represents the combination of optimization options that involves respectively options number 2, 3, 7, 8, 9, and 10 in optimization. Since the binary representation used to represent the chromosome, the uniform (regular) crossover with either single point or more can be implemented depends on how many options compiler provides; while the mutation can be done by simply flipping the value in the single mutated location. In this case, the **fitness function** is defined by summing each $RIP_B(F_i = 0)$ (adopted from equation 3) in each C_j, and the formal definition is shown in equation 8.

$$Fitness(C_j) = \sum_{i=1}^{n} RIP_B(F_i | F_i = 0) \tag{8}$$

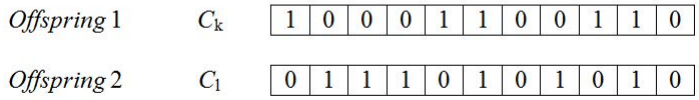
where C_j is one of the combination in population, and n is the number of optimization options $\{F_1, F_2, \ldots, F_n\}$.

Equation 3 says that, the optimization option of F_i has a negative effect when the value of $RIP_B(F_i = 0) < 0$. So that, according to the selection criteria, only the combinations with higher value of fitness function (i.e., fitter) will be considered as a parent candidate for next generation (iteration).

Having mentioned at the previous discussion, and by considering the representation of chromosome; the uniform (or regular) crossover with either single point or more would be implemented in the reproducing process of individual for the next generation. The process of crossover is illustrated in Fig. 1 and Fig. 2.

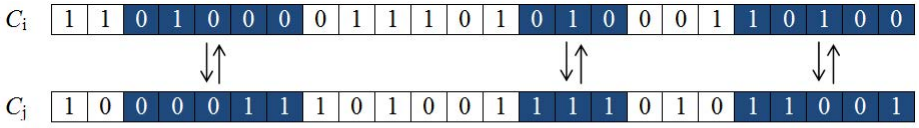


(a) The process of Regular Crossover Operation with single-point

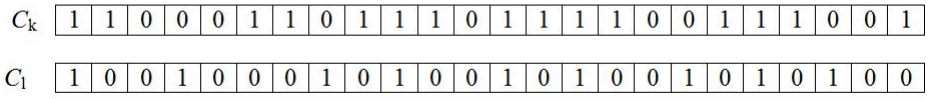


(b) The result of Regular Crossover Operation

Fig. 1. The operation of Regular Crossover with two-point



(a) The process of Regular Crossover Operation with five-point



(b) The result of Regular Crossover Operation with five-point

Fig. 2. The operation of Regular Crossover with five-point

It is unlike the crossover operator which is binary, mutation is unary. First, the mutated location is determined randomly, and the value of that location is then replaced by only flipping the value from '0' to '1' and vice versa. The process of mutation is shown in Fig. 3.

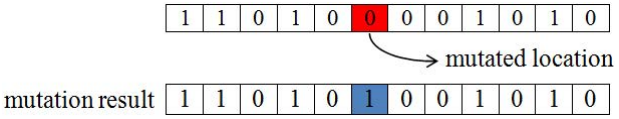


Fig. 3. The operation of mutation operator

The pseudo code of HE algorithm is as follows.

1. Determine genetics's parameters, i.e., the size of population, the probability of crossover, and the probability of mutation respectively m, p_c, and p_m,
2. Generate the initial population (collection of the combination of optimization options) randomly, $P = \{C_1, \ldots, C_m\}$, with $C_i = \{F_1 = 0 \text{ or } 1, \ldots, F_n = 0 \text{ or } 1\}$,
3. Compute the value of fitness of each chromosome C_j using equation 8,
4. Based on fitness values computed in the previous step, and certain selection method (for instance Roulette wheel selection), select chromosome as parent candidates,
5. Check the termination condition, if the condition is false, then do step (6); otherwise STOP.
6. Crossover the parent by considering the value of p_c, to yield new chromosomes (*offspring*) as many as the population size for the next generation, go to steps (3) - (5).

Note that the termination condition could be either the determined number of generations (iterations) or some determined value of threshold as an indicator of its convergence.

4 Conclusion

In accordance to the way the algorithm find the best combination, HE algorithm is only relevant to be compared with ES, BE and OSE algorithms. The following are some remarks about that comparison.

- As the name implies, ES algorithm finds the best combination of the compiler's optimization options by exhaustively checking all possible ones.
- The result of BE algorithm is an optimal combination of the compiler's optimization options obtained by removing optimization option with the most negative RIP iteratively.
- OSE algorithm build the combination of compiler optimization options starting from the set with single default optimization combination, then iteratively the set is updated by performing the union operation.
- HE algorithm finds the optimal solution (combination of compiler's optimization options) starting from the initial population contains a number of combinations of compiler's optimization options which were initially chosen

in random manner. Then, each chromosomes (representation of combinations) in the population were evaluated using fitness function to determine fitter chromosomes to be chosen for the next generation. This process was performed iteratively until some determined condition satisfied, and the best combination is obtained.

References

1. Hedayat, A., Sloane, N., Stufken, J.: Orthogonal Arrays: Theory and Applications. Springer (1999)
2. Chow, K., Wu, Y.: Feedback-directed selection and characterization of compiler optimizations. In: Second Workshop on Feedback Directed Optimizations, Israel (November 1999)
3. Box, G.E.P., Hunter, W.G., Hunter, J.S.: Statistics for Experimenters: an introduction to design, data analysis, and model building. John Wiley and Sons (1978)
4. Chow, K., Wu, Y.: Feedback-directed selection and characterization of compiler optimizations. In: Second workshop of Feedback-directed Optimizations, Israel (November 1999)
5. Nadia, N., Ajith, A., Luiza de Macedo, M.: Genetic Systems Programming - Theory and Experience. Springer (2006)
6. Sloane, N.J.A.: A Library of Orthogonal Arrays, `http://www.research.att.com/njas/oadir/`
7. Kulkarni, P., Hines, S., Hiser, J., Whalley, D., Davidson, J., Jones, D.: Fast Searches for Effective Optimization Phase Sequences. In: PLDI 2004: Proceeding of the ACM SIGPLAN 2004 Conference of Programming Language Design and Implementation, pp. 171–182. ACM Press, New York (2004)
8. Pinkers, R.P.J., Knijnenburg, P.M.W., Haneda, M., Wijshoff, H.A.G.: Statistical selection of compiler optimizations. In: The IEEE Computer Societies 12th Annual International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems (MAS - COTS 2004), Volendam, The Netherlands, pp. 494–501 (October 2004)
9. Triantafillis, S., Vacharajani, M., Vacharajani, N., August, D.I.: Compiler Optimization-space Exploration. In: Proceedings of the International Symposium on Code generation and Optimization, pp. 204–215 (2003)
10. Kisuki, T., Knijnenburg, P.M.W., O'Boyle, M.F.P., Bodin, F., Wijshoff, H.A.G.: A Feasibility Study in Iterative Compilation. In: Fukuda, A., Joe, K., Polychronopoulos, C.D. (eds.) ISHPC 1999. LNCS, vol. 1615, pp. 121–132. Springer, Heidelberg (1999)
11. Pan, Z., Eigenmann, R.: Compiler Optimization Orchestration for peak performance. Technical Report TR-ECE-04-01. School of Electrical and Computer Engineering, Purdue University (2004)
12. Pan, Z., Eigenmann, R.: Rating Compiler Optimizations for automatic performance tuning. In: SC 2004: High Performance Computing, Networking and Storage Conference, 10 pages (November 2004)
13. Pan, Z., Eigenmann, R.: Rating Compiler Optimizations for Automatic Performance Tuning. IEEE (2004)

On Efficient Processing of Complicated Cloaked Region for Location Privacy Aware Nearest-Neighbor Queries

Chan Nam Ngo and Tran Khanh Dang

Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology, VNUHCM, Vietnam
khanh@cse.hcmut.edu.vn

Abstract. The development of location-based services has brought not only conveniences to users' daily life but also great concerns about users' location privacy. Thus, privacy aware query processing that handles cloaked regions has become an important part in preserving user privacy. However, the state-of-the-art private query processors only focus on handling rectangular cloaked regions, while lacking an efficient and scalable algorithm for other complicated cloaked region shapes, such as polygon and circle. Motivated by that issue, we introduce a new location privacy aware nearest-neighbor query processor that provides efficient processing of complicated polygonal and circular cloaked regions, by proposing the Vertices Reduction Paradigm and the Group Execution Agent. In the Vertices Reduction Paradigm, we also provide a new tuning parameter to achieve trade-off between answer optimality and system scalability. Finally, experimental results show that our new query processing algorithm outperforms previous works in terms of both processing time and system scalability.

Keywords: Location-based service, database security and integrity, user privacy, nearest-neighbor query, complicated cloaked region, group execution.

1 Introduction

To preserve the LBS user's location privacy, the most trivial method is to remove the direct private information such as identity (e.g., SSID). However, other private information, such as position and time, can also be used to violate the user's location privacy [3]. In preventing that, the researchers have introduced the Location Anonymizer [3]. It acts as a middle layer between the user and the LBS Provider to reduce the location information quality in the LBS request. The quality reduction is performed by the obfuscation algorithm which transforms the location to be more general (i.e., from a point to a set of points [9], a rectilinear region [10-15], or a circular region [6], etc.). The request is then sent to the LBS Provider to process without the provider knowing the user's exact location. Due to the reduction in location quality, the LBS Provider returns the result as a candidate set that contains the exact answer. Later, this candidate set can be filtered by the Location Anonymizer to receive the request's exact answer for the LBS user. Consequently, to be able to process those requests, the LBS Provider's Query Processor must deal with the cloaked region rather than the

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 101–110, 2013.

exact location. In this paper, we propose a new Privacy Aware Nearest-Neighbor (NN) Query Processor that extends Casper* [8]. Our Query Processor can be embedded inside the untrusted location-based database server [8], or plugged into an untrusted application middleware [3]. The Privacy Aware Query Processor is completely independent of the location-based database server in the LBS Provider and underlying obfuscation algorithms in the Location Anonymizer. Moreover, it also supports various cloaked region shapes, which allows more than one single obfuscation algorithm to be employed in the Location Anonymizer [3]. In addition, we introduce a new tuning parameter to achieve trade-off between candidate set size and query processing time. Finally, we propose an additional component for the Location Anonymizer, the Group Execution Agent, to strongly enhance the whole system's scalability. Our contributions in this paper can be summarized as follows:

- We introduce a new Privacy Aware NN Query Processor. With its Vertices Reduction Paradigm (VRP), complicated polygonal and circular cloaked regions are handled efficiently. In addition, the performance can be tuned through a new parameter to achieve trade-off between candidate set size and query processing time.
- We propose an addition component for the Location Anonymizer, the Group Execution Agent (GEA), to strongly enhance the whole system's scalability.
- We provide experimental evidence that our Privacy Aware Query Processor outperforms previous ones in terms of both processing time and system scalability.

The rest of the paper is organized as follows. In section 2 and 3, we highlight the related works and briefly review the Casper* Privacy Aware Query Processor. The proposed Vertices Reduction Paradigm and Group Execution Agent are discussed in section 4 and 5 respectively. Then we present our extensive experimental evaluations in section 6. Lastly, section 7 will finalize the paper with conclusion and future works.

2 Related Works

In general, Privacy Aware Spatial Data are classified into Public Data (exact location, *public object* such as Point-Of-Interest or public forces' (police officers) location) and Private Data (cloaked region). Based on that classification, a Privacy Aware Query Processor must have the ability to process four types of query [8], includes one exact query type (exact traditional spatial query) and three privacy aware ones: (1) Private Query over Public Data, e.g., "*Where's the nearest restaurant?*", in which the user's location is a *private object* while the restaurant's (answer) is public, (2) Public Query over Private Data, i.e., "*Who have been around this car during 6AM to 8AM?*", the user is a policeman whose location is a *public object*, he is looking for suspects whose locations are *private objects*, (3) Private Query over Private Data, e.g., "*Where's my nearest friends?*", the user is asking for their nearest friends in friends finder service, in which both locations are *private objects*. Recently, Duckham et al. have proposed an obfuscation algorithm [9] that transforms an exact user location to a *set of points* in a road network based on the concepts of *inaccuracy* and *imprecision*. They also provide a NN query processing algorithm. The idea is that the user will first send the

whole *set of points* to the server, the server will send back a *candidate set of NNs*. Based on that candidate set, the user can either choose to reveal more information in the next request for more accurate result or terminate the process if satisfied with the candidate set of NNs. The other works in [4-5] respectively propose algorithms to deal with *circular* and *rectilinear* cloaked region, those works find the *minimal set of NNs*. In a different approach, Casper* only computes a *superset* of the *minimal set of NNs* that contains the exact NN, in order to achieve trade-off between query processing time and candidate set size for system scalability [8]. In addition, Casper* also supports two more query types: Private and Public Query over Private Data.

Among previous works, only Casper* supports Query over Private Data, while the others either only support Query over Public Data [9] or lack the trade-off for system scalability [4-5]. However, Casper* is only efficient in dealing with rectangular regions. While it can handle polygonal cloaked regions, the application into these cases needs evaluations and modifications. Moreover, in case of systems like OPM [3], the Query Processor must have the ability to deal with various kinds of cloaked region because the system supports more than one single obfuscation algorithm. Motivated by those problems, our proposed Privacy Aware NN Query Processor offers the ability to efficiently handle complicated polygonal and circular cloaked regions with its Vertices Reduction Paradigm and a new tuning parameter for system scalability. Furthermore, we provide an addition component for the Location Anonymizer, the Group Execution Agent, to strongly enhance the whole system's scalability.

3 The Casper* Privacy Aware Query Processor

In this section, let us briefly review the Casper* algorithm, start with its terms and notations. For each vertex v_i of the cloaked region A, its NN is called a *filter*, denoted as t_i if that NN is a public object (*public NN*) (Fig. 1b, t_1, t_2 of v_1, v_2). In case the NN is private, it is denoted as At_i. A private object is considered as *private NN* if it has the minimum distance from its cloaked region's furthest corner to v_i (Fig. 1d, At_1). The distance between a vertex and its filter is denoted as $dist(v_i, t_i)$ (public NN) or $min\text{-}max\text{-}dist(v_i, At_i)$ (private NN). For each edge e_{ij} formed by adjacent vertices v_i, v_j, a *split-point* s_{ij} is the intersection point of e_{ij} and the perpendicular bisector of the line segment t_it_j (Fig. 1b, s_{12}). For the purpose of the Casper* NN algorithm [8], given a cloaked region A, it is to find all the NNs of all the points (1) inside A and (2) on its edges. The algorithm can be outlined in the three following steps below.

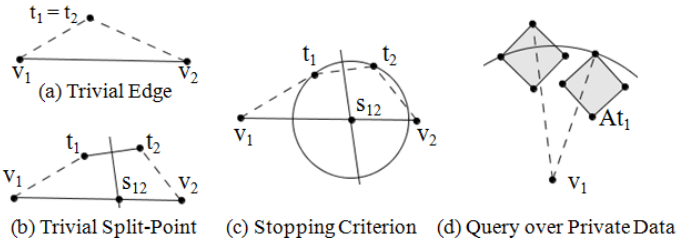


Fig. 1. The Casper* Algorithm

- **STEP 1** (*Filter Selection*): We find the filters for all of cloaked region A's vertices.
- **STEP 2** (*Range Selection*): For each edge e_{ij} of cloaked region A, by comparing v_i, v_j's filters t_i and t_j, we consider four possibilities to find *candidate NNs* and *range searches* that contain the candidate NNs.
  - ***Trivial Edge Condition***: If $t_i = t_j$ (Fig. 1a, $t_1 = t_2$), t_i (t_j) is the NN to all the points on e_{ij}, so we add t_i (t_j) into the *candidate set*.
  - ***Trivial Split-Point Condition***: In this case, $t_i \neq t_j$, but *split-point* s_{ij} of e_{ij} takes t_i, t_j as its NNs (Fig. 1b). This means t_i and t_j are the NNs to the all points on v_is_{ij} and $s_{ij}v_j$ respectively. So we add t_i, t_j into the *candidate set*.
  - ***Recursive Refinement Condition***: If two conditions above fail, we will consider to split the edge e_{ij} into v_is_{ij} and $s_{ij}v_j$, then we apply STEP 2 to them recursively. A parameter *refine* is used to control the recursive calls for each edge, it can be adjusted between 0 and ∞ initially in the system. For each recursive call, *refine* will be decreased by 1, and when it reaches 0, we will stop processing that edge. In this case, $refine > 0$, we decrease it by 1 and process v_is_{ij} and $s_{ij}v_j$ recursively.
  - ***Stopping Criterion Condition***: When *refine* reaches 0, we add the circle centered at s_{ij} of a radius $dist(s_{ij}, t_i)$ as a range query into the *range queries set R* and stop processing current edge (Fig. 1c).
- **STEP 3** (*Range Search)*: we execute all the range queries in R, and add the objects into the *candidate set*. As a result, the *candidate set* contains NNs for all the points (1) inside cloaked region A and (2) on its edges. After that, the *candidate set* will be sent back to the Location Anonymizer to filter the exact NN for the user.

In Query over Private Data, STEP 2 is similar to Query over Public Data, with some modifications. Instead of adding At_i directly into the *candidate set*, we will have to add a circle centered at v_i of a radius $min\text{-}max\text{-}dist(v_i, At_i)$ as a range query into the *range queries set R* (Fig. 1d). The same behavior is applied to v_j and s_{ij} of edge e_{ij}.

4 Vertices Reduction Paradigm

Although Casper* can deal with polygonal cloaked region A that has n vertices (n-*gon*), its runtime significantly depends on A's number of vertices (STEP 1) and edges (STEP 2). As shown in Fig. 2a's formula 1 and 2, to process an n-*gon*, Casper* suffers from two aspects. (1) The processing time of STEP 1 increases because it has to find more filters ($4Qt_4 \leq nQt_n$). Besides, the calculation for $min\text{-}max\text{-}dist(v_i, At_i)$ also increase the range query runtime for the n-*gon* ($Qt_4 \leq Qt_n$). (2) The processing time of STEP 2 increases as it has to process more edges ($4(2^{refine} - 1)Qt_4 \leq n(2^{refine} - 1)Qt_n$).

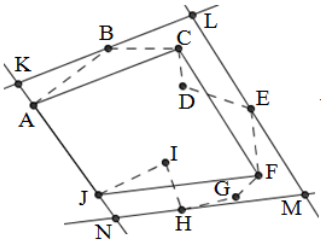
To ease that problem, we introduce the *Vertices Reduction Paradigm* (VRP), in which we simplify the polygon so that it has as less vertices as possible before processing it with the Casper* algorithm. For that purpose, the *Ramer–Douglas–Peucker* (RDP) [1] algorithm is employed. For each private object (n-*gon*) in the database, we maintain a *vertices reduced version* (VRV, m-*gon*, $m < n$) of that private object. The VRV is generated by the RDP algorithm and it will be stored inside the database until invalidated. For NN query processing, we use the VRVs instead of original ones to reduce processing time ($m \leq n$ and $Qt_m \leq Qt_n$, as depicted in formula 3 of Fig. 2a).

The purpose of the RDP [1] algorithm, given an *n-gon* (*ABCDEFGHIJ* in Fig. 2b), is to find a subset of fewer vertices from the *n-gon*'s list of vertices. That subset of vertices forms an *m-gon* that is simpler but similar to the original *n-gon* ($m < n$). The inputs are the *n-gon*'s list of vertices and the *distance dimension* $\varepsilon > 0$. First, we find the vertex that is furthest from the line segment with the first and last vertices as end points. If that furthest vertex is closer than ε to the line segment, any other vertices can be discarded. Otherwise, given the *index* of that vertex, we divide the list of vertices into two: *[1..index]* and *[index..end]*. The two lists are then processed with the algorithm recursively. The output is an *m-gon*'s list of vertices (Fig. 2b, *ACFJ*).

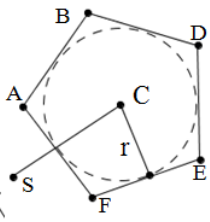
In next subsections, we will discuss two different approaches to utilize the VRVs. The first one is to use the VRV directly. The second one, which is better than the first, is to use the bounding polygon of the VRV. In both approaches, the RDP algorithm's overhead in computing the VRVs is insignificant compared to the total processing time of the query. As depicted in Fig. 2b, the dotted polygon *ABCDEFGHIJ* is the original *n-gon* while *ACFJ* and *KLMN* is the *m-gon* of the first and second approach respectively. For a circular region, the VRV is its bounding polygon (Fig. 2c, *ABDEF*) and we use the distance from another vertex to its center plus the radius as *min-max-dist* of it and that vertex in private NN search ($SC+r$ in Fig. 2c).

Rectangle	$4Qt_4+4(2^{refine}-1)Qt_4$	(1)
Polygon	$nQt_n+n(2^{refine}-1)Qt_n$	(2)
VRP	$mQt_{m'}+m(2^{refine}-1)Qt_m$	(3)
Number of Vertices (4, m, n)	$4 \le m \le n$	
Range Query Runtime (Qt_4, Qt_m, Qt_n)	$Qt_4 \le Qt_m \le Qt_n$	

(a) Computational Cost



(b) Polygonal VRV



(c) Circular VRV

Fig. 2. The Vertices Reduction Paradigm

4.1 The Direct Vertices Reduction Paradigm Approach

In this approach, by using the *m-gons* as the cloaked regions of the query and the private objects, we reduce the query processing time in STEP 1 and STEP 2 of the Casper* algorithm (Fig. 2a, formula 3). However, since we use an approximate version of the original cloaked region, we need some modifications in STEP 2 to search for NNs of the parts of *n-gon* that are outside the *m-gon* (e.g., *ABC* in Fig. 2b). During the RDP's recursive divisions, for each simplified edge, we maintain the distance of the furthest vertex that is not inside the *m-gon* (*A, B, E* and *H* in Fig. 2b). The list's size is exactly *m*. We denote those distances as *d* (Fig. 2b, distance from *H* to *FJ*). The modifications make use of the *distance d* and only apply to the simplified edges that the discarded vertices are not all inside the *m-gon*, e.g. *AC*, *CF* and *FJ* in Fig. 2b.

- **Modifications for Query over Public Data**
  - ***Trivial Edge and Split-Point Condition***: using the corresponding *distance d* maintained above, we add two range queries centered at v_i, v_j of radii $dist(v_i, t_i) + d$, $dist(v_j, t_j) + d$ into the *range queries set R* (Fig. 3a). For Trivial Split-Point Condition, we add one more range query centered at s_{ij} of a radius $dist(s_{ij}, t_i) + d$ into *R*. As shown in Fig. 3c, the NN *E'* of any point *H* on *BC* (*C* is a discarded vertex outside the *m-gon*) must be inside the hatched circle centered at *H* of radius *HE* (*HE'* ≤ *HE*), which is always inside the two bold circles created by the enlarged (+*d*) range queries. It is also the same for any points in *ABC*.
  - ***Stopping Criterion Condition***: similarly, we increase the range query's radius by *d* to ensure including the NNs of the outside parts of the original *n-gon*.
- **Modifications for Query over Private Data**: because the private objects are also simplified, we will *increase the search radius by d + ε* for not missing them as candidate NNs (depicted in Fig. 3b, the range query *(+d+ ε)* reaches the simplified edge *AB* of another private object while the range query *(+d)* does not).

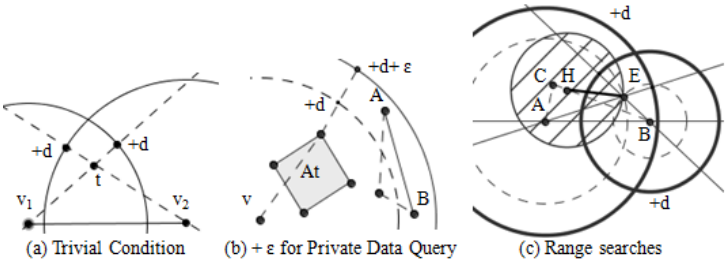


Fig. 3. Modifications in Vertices Reduction Paradigm

4.2 The Bounding Vertices Reduction Paradigm Approach

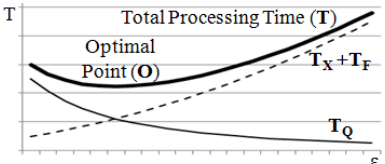
One characteristic of the *m-gon* generated by the RDP algorithm is that it may not contain the original *n-gon*. In this approach, we want to ensure the *m-gon* contains the original one. During the RDP's recursive divisions, for each simplified edge (*m* edges), we maintain the furthest vertex that is not inside the *m-gon* (*A, B, E* and *H* in Fig. 2b). After that, we calculate the *m* lines that are parallel to the respective edges of the *m-gon* and through the respective furthest vertices in the list (e.g., *KL, LM, MN* and *NK* in Fig. 2b). The intersection of those lines forms a new *m-gon* that contains the original *n-gon* inside it (Fig. 2b's *KLMN*). Therefore, the candidate set of the simplified *m-gon* is a superset of the original *n-gon* without directly modifying Casper*.

Although the first approach reduces the query processing time much, it suffers from the moderate increase of the *candidate set* size. Differently, the second approach achieves both better *candidate set* size and query processing time than the first one. Firstly, we can add the filters directly into the *candidate set* without the risk of missing the exact NN because the *m-gon* contains the original *n-gon* (no outside parts). Secondly, although the range query's radius is indirectly enlarged through the enlargement of the original *n-gons* to the bounding *m-gons*, it is kept minimum (+*d*+*d*, an indirect +*d* of the cloaked region

and another $+d$ of the private object). Thus the number of results for each range query is also reduced. Furthermore, the reduction in number of range queries also leads to a slight reduction of processing time.

4.3 The Distance Dimension ε as VRP Tuning Parameter

The *Total Processing Time* (T) of a query consists of three components. (1) The *Query Processing Time* (T_Q), which is for the Query Processor to compute the candidate set. (2) The *Data Transmission Time* (T_X) which is for the candidate set to be transmitted to the Location Anonymizer for NNs filtration. (3) The *Answers Filtration Time* (T_F) which is for the candidate set to be filtered for exact NN of the query request. T_Q is monotonically decreasing with the decrease of number of vertices, while T_X and T_F are monotonically decreasing with the decrease of candidate set size. Thus, we can utilize the distance dimension ε as a tuning parameter for VRP since it affects the number of vertices in the VRV and the search radius of range queries in the range query set R. We will consider 2 cases in respect to the ε value: (1) $T_Q > T_X+T_F$. Initially, the ε is too small that query processing takes too much time. To resolve this, we must increase ε. (2) $T_Q < T_X+T_F$. This indicates the candidate set size is too large that (T_X+T_F) is longer than T_Q. We have to decrease ε. Thus, in order to find an optimal value of ε for the best T, we increase ε until it reaches the optimal point O (Fig. 4a).



(a) Tuning with ε

VRP	$N(T_Q + T_X) + NT_F$	(1)
GEA	$K(T_Q + T'_X) + NT'_F$	(2)
Number of Queries (N, K)		$K \leq N$
Average Transimision Time (T_X, T'_X)		$KT'_X \leq NT_X$

(b) Group Execution Agent

Fig. 4. System Scalability

5 Group Execution Agent

As shown in Fig. 5, there are many queries with adjacent and overlapped regions at a time (the dotted regions), or even better, a query's region is contained inside another's. Obviously, such queries share a part of or the whole candidate set. To take advantage of that, we propose the Group Execution (GE) algorithm for the Location Anonymizer's additional component, the Group Execution Agent (GEA). The GEA will group as many queries as possible for one query execution before sending them to the Query Processor (N queries into K query groups, $K \leq N$, i.e., 9 queries to 3 groups in Fig. 5, the bold $G_{1,2,3}$ are used as cloaked regions in NN queries).

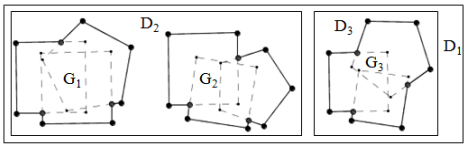


Fig. 5. Group Execution

Algorithm. Group Execution

function GroupExecution(**list of query regions, system-defined max$_A$**)
while true **and** list size > 1 **do**
  for each region r_i, find another region r_j that has least enlargement if r_i is grouped with r_j, the new region's area $a \leq max_A$, and add into *list*(r_i,r_j,a)
  break if *list*(r_i,r_j,a) is empty
  sort *list*(r_i,r_j,a) by a ascending
  for each r_i **in** *list*(r_i,r_j,a)
    if r_i already grouped in this loop
      if r_j already grouped in this loop **continue**
      else *groupedRegions* = *groupedRegions* U {r_j}
    else *groupedRegions* = *groupedRegions* U GroupedRegionOf(r_i, r_j)
  regions = *groupedRegions*
return *regions (maximum number of regions grouped, minimum area each group)*

The algorithm is outlined in the pseudo code above. Its purpose, given a *list of query regions* (of size N) and a parameter max_A, is to group the regions in the list into K *grouped regions* ($K \leq N$) of which areas are smaller than max_A. Then the queries are also put into *query groups* accordingly. The *grouped regions* are used as the cloaked regions of those *query groups* in NN query processing. The *query group*'s candidate set is a superset of the candidate set of the queries in the group, so the GEA does not miss any exact NN of those queries. The system benefits from the GEA as shown in Fig. 4b. (1) The query processing time for each *query group* is the same as a single query (T_Q) because we only execute the *query group* once with the *grouped region* as input. Thus, the sum of all queries' processing time decreases ($KT_Q \leq NT_Q$). This also leads to the decrease of average query processing time. (2) The *query group*'s candidate set size increases because it is a superset of the candidate set of those queries in the group, but the average transmission time decreases as we only transmit the common candidates once (as $KT'_X \leq NT_X$). The average filtration time increases ($T'_F \geq T_F$), but it is minor in comparison to the benefits above. Furthermore, for optimization, the algorithm's input list must satisfies two conditions: (1) the list's regions are adjacent to each other for easier grouping, (2) the list size is small enough to avoid scalability problem because the algorithm's complexity is $O(n^2)$. To find those suitable lists, we maintain an R*-Tree [2] in the Location Anonymizer. When a query is sent to the Anonymizer, its cloaked region is indexed by the R*-Tree. By finding the R*-Tree's nodes of which the directory rectangle's area are smaller than a predefined area value $kmax_A$, we will get the suitable lists from those nodes' regions. In Fig. 5, we find two suitable lists from the nodes D_2 and D_3's regions (D_1's area > $kmax_A$). Later, the algorithm returns *grouped regions* G_1, G_2 and G_3, which reduces the number of queries from 9 to 3. In fact, the GEA's speedup is dependent of how much overlapped the regions are. The worst case could be that we cannot group any query but still have the overhead of the R*-Tree and the GE algorithm. However, in most cases, when the number of queries is large enough, the GEA does strongly reduce the system's average query processing and transmission time and improve the system scalability.

6 Experimental Evaluations

We evaluate both two VRP approaches and the GE algorithm for the Private Query over Public and Private Data. The algorithms are evaluated with respect to the tuning parameter ε. For all two types of private query, we compare our algorithms with the Casper*, the performance evaluations are in terms of total processing time and candidate set size. We conduct all experiments with 100K private data and 200K public data. The polygonal cloaked regions are generated by Bob-Tree [14-15], and the circular ones are generated by the works in [6]. The polygonal cloaked regions' number of vertices range from 20 to 30, while the value of ε is varied in the range of 10% to 50% of the Bob-Tree's grid edge size. For GEA, the number of queries is 10K and the parameter max_A and $kmax_A$ are 30 and 100 times of the above grid's area respectively.

The charts in Fig. 6 show our experimental results. As shown in the processing time charts, the VRPs perform significant improvements compared to Casper*. When ε increases, the processing time of VRPs and GEA decrease while Casper*'s remains constant. Because the larger the ε value is, the larger reduction we can achieve, leads to the larger reduction in query processing time, especially in Private Query over Private Data. At the largest ε (50%), the VRPs reduce the total processing time by 98% for Private Query over Private Data (Fig. 6b) with the standard variation at only 92ms (10% of average total processing time). However, the candidate set size increases moderately (Direct VRP) and slightly (Bounding VRP). Lastly, with the additional component GEA, the total processing time and candidate set size are reduced at best (ε = 50%) by 66% and 33% respectively in comparison to Bounding VRP, the best VRP approach. This helps ease the increase in candidate set size of VRP.

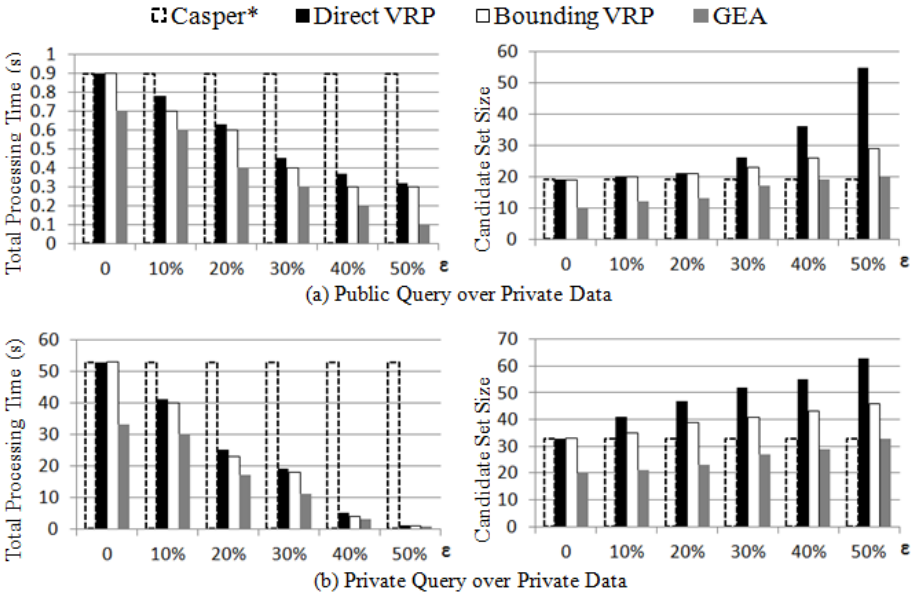


Fig. 6. Experimental Evaluations

7 Conclusion and Future Works

In this paper, we introduce a new Privacy Aware Query Processor that extends Casper*. With the new Query Processor's Vertices Reduction Paradigm and its tuning parameter ε, complicated polygonal [12-15] and circular [6] cloaked regions are handled efficiently. The main idea is that we employ the Ramer–Douglas–Peucker algorithm to simplify the region's polygon before processing it. Furthermore, we propose the Group Execution Agent to strongly enhance the system scalability. Experimental results show that our works outperform Casper* in dealing with such kinds of region above. For future, we will consider supporting k nearest-neighbor query, continuous query [7-8] and trajectory privacy [16] in our Privacy Aware Query Processor.

References

1. Douglas, D.H., Peucker, T.K.: Algorithms for the Reduction of The number of Points required to represent a Digitized Line Or its Caricature. In: Cartographica: The Intl. Journal for Geographic Information and Geovisualization, pp. 112–122. Univ. of Toronto Press (1973)
2. Beckman, N., Kriegel, H., Schneider, R., Seeger, B.: The R*-tree: an efficient and robust access method for points and rectangles. In: SIGMOD, pp. 322–331 (1990)
3. Dang, T.K., Phan, T.N., Ngo, C.N., Ngo, N.N.M.: An open design privacy-enhancing platform supporting location-based applications. In: ICUIMC, pp. 1–10 (2012)
4. Hu, H., Lee, D.L.: Range Nearest-Neighbor Query. IEEE TKDE 18(1), 78–91 (2006)
5. Kalnis, P., Ghinita, G., Mouratidis, K., Papadias, D.: Preventing Location-Based Identity Inference in Anonymous Spatial Queries. IEEE TKDE 19(12), 1719–1733 (2007)
6. Ardagna, C.A., Cremonini, M., Damiani, E., Vimercati, S.D.C., Samarati, P.: Location privacy protection through obfuscation-based techniques. In: DBSec, pp. 47–60 (2007)
7. Truong, Q.C., Truong, T.A., Dang, T.K.: The Memorizing Algorithm: Protecting User Privacy in Location-Based Services using Historical Services Information. IJMCMC 2(4), 65–86 (2010)
8. Chow, C., Mokbel, M.F., Aref, W.G.: Casper*: Query processing for location services without compromising privacy. ACM TODS, 1–48 (2009)
9. Duckham, M., Kulik, L.: A formal model of obfuscation and negotiation for location privacy. In: PerCom, pp. 243–251 (2005)
10. Dang, T.K., Truong, T.A.: Anonymizing but Deteriorating Location Databases. POLIBITS 46, 73–81 (2012)
11. Truong, A.T., Dang, T.K., Küng, J.: On Guaranteeing k-Anonymity in Location Databases. In: Hameurlain, A., Liddle, S.W., Schewe, K.-D., Zhou, X. (eds.) DEXA 2011, Part I. LNCS, vol. 6860, pp. 280–287. Springer, Heidelberg (2011)
12. Damiani, M.L., Bertino, E., Silvestri, C.: The PROBE Framework for the Personalized Cloaking of Private Locations. TDP 3(2), 123–148 (2010)
13. Le, T.T.B., Dang, T.K.: Semantic-Aware Obfuscation for Location Privacy at Database Level. In: Mustofa, K., Neunold, E., Tjoa, A M., Weippl, E., You, I. (eds.) ICT-EurAsia 2013. LNCS, vol. 7804, pp. 111–120. Springer, Heidelberg (2013)
14. To, Q.C., Dang, T.K., Küng, J.: B^{ob}-Tree: An Efficient B^{+}-Tree Based Index Structure for Geographic-Aware Obfuscation. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011. LNCS (LNAI), vol. 6591, pp. 109–118. Springer, Heidelberg (2011)
15. To, Q.C., Dang, T.K., Küng, J.: A Hilbert-based Framework for Preserving Privacy in Location-based Services. In: IJIIDS (2013)
16. Phan, T.N., Dang, T.K.: A Novel Trajectory Privacy-Preserving Future Time Index Structure in Moving Object Databases. In: ICCCI, pp. 124–134 (2012)

Semantic-Aware Obfuscation for Location Privacy at Database Level

Thu Thi Bao Le and Tran Khanh Dang

Faculty of Computer Science & Engineering, HCMUT
Ho Chi Minh City, Vietnam
{thule,khanh}@cse.hcmut.edu.vn

Abstract. Although many techniques have been proposed to deal with location privacy problem, which is one of popular research issues in location based services, some limitations still remain and hence they cannot be applied to the real world. One of the most typical proposed techniques is obfuscation that preserves location privacy by degrading the quality of user's location information. But the less exact information, the more secure and the less effective services can be supported. Thus the goal of obfuscated techniques is balancing between privacy and quality of services. However, most of obfuscated techniques are separated from database level, leading to other security and performance issues. In this paper, we introduce a new approach to protect location privacy at database level, called Semantic B^{ob}-tree, an index structure that is based on B^{dual}-tree and B^{ob}-tree and contains semantic aware information in its nodes. It can achieve high level of privacy and keep up the quality of services.

Keywords: Privacy, Security, Location based services, Obfuscation.

1 Introduction

Location Based Service (LBS) is a concept that denotes applications supplying utilities to users by using their geographic location (i.e., spatial coordinates) [9]. In recent years, with the development of technologies including mobile devices, modern positioning techniques such as Global Positioning System (GPS), internet (like wireless, 3G), Location Based Services have become more and more popular and have brought many utilities for human life.

However, when using these applications, the users will be asked to send their sensitive information to the service provider, such as location, identity, name, etc. Revealing this information may cause some dangerous. The attackers may know the personal information that the users want to keep secret, such as true position at current time, or the daily schedule. Therefore, privacy in LBS is a serious problem and a lot of research has been done in this field. Note that the more accurate information is, the more easily the privacy is revealed. Therefore the main problem is that how to balance between the quality of services and the privacy.

To solve this privacy-preserving problem, many techniques are suggested, such as k-anonymity based approaches [8, 17, 18], policy based approaches [14], etc. Among

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 111–120, 2013.

them, the most popular one is the obfuscation-based approach. The general idea of this technique is to degrade the quality of user's location information but still allow them to use services with acceptable quality [1, 4]. However, this technique has some limitations. Most of the algorithms that belong to the obfuscation techniques are geometry-based. They do not consider the geographic feature inside the obfuscated region. Based on the knowledge about the geography of the region, the adversary can increase the probability of inferring the exact user's location [3]. Another limitation is that these algorithms are separated from the database level. This makes the algorithms go through two phases: First, retrieving the accurate location of user at the database level, and then obfuscating this information at the algorithm level. So the time cost is increased and security is harder to obtain [6, 7, 12, 13].

Motivated by these issues, we will propose a new semantic-aware obfuscation technique, called Semantic B^{ob}-tree, where semantic of regions are taken into each user's account. This technique is embedded into a geo-database in order to reduce the overall cost.

The rest of the paper is organized as follows. We review related work in section 2. Next, in section 3, we propose a new idea for obfuscation, called Semantic B^{ob}-tree. Following them, the evaluation is discussed in section 4 before we make our conclusion and future work in section 5.

2 Related Work

2.1 Location Obfuscation

There are many approaches to location privacy problem. Among the most popular techniques to protect user's location, the obfuscation gained much interest [1, 2, 10, 19]. Location obfuscation is the process of degrading the quality of information about a person's location, with the aim of protecting that person's location privacy [1, 4]. In [1, 11], the authors propose obfuscation techniques by enlarging, shifting, and reducing the area containing real user's location.

However, these obfuscation techniques just deal with geometry of the obfuscated region, not concern about what are included inside. The adversary with geographical knowledge may infer sensitive location information from obfuscated location generated by geometric based techniques. For example, if an adversary knows that a person is in a region just contains a hospital and a lake. But that person cannot be in the lake (assume no one can be in the lake), so the adversary may easily infer that he/she is in the hospital. We call this privacy attack as spatial knowledge attack.

2.2 Semantic-Aware Obfuscation

Because geometry-based techniques cannot protect location privacy if the adversary has geographical knowledge about obfuscated region, the semantic-aware obfuscation technique has been proposed in [3]. This technique considers sensitive feature types inside the obfuscated region. However, this technique does not concern about how big the area of the obfuscated region is. In some LBS applications, the requirement is that

the area of the obfuscated region must be big enough to protect user's location privacy and must be small enough to ensure the quality of services.

2.3 B^{ob}-Tree

B^{ob}-tree [6, 7, 12] is based on B^{dual}-tree [5] and contains geographic-aware information on its node. The process of calculating the obfuscated region can be done in only one phase: traversing the index structure to retrieve the appropriate obfuscated region that contains user's location. This one-phase process can reduce the processing time. However, in this approach, the region is divided into just two geographic features: approachable and unapproachable parts. The criteria of this division are completely based on geographic features. For example the lakes, the mountains are unapproachable parts, while the parks, the schools are approachable parts. But the user may need more semantic for his/her particular case, such as if the user is a common citizen, the military zone may be unapproachable, but if the user is a soldier, it is an approachable part. In the other words, the adversary may infer sensitive location information by using information about user and geographical knowledge. Moreover, a region can be sensitive for someone but non sensitive for another, or a place is in high sensitivity for someone but is in low sensitivity for another.

Motivated by this, we combine two ideas of semantic aware obfuscation and B^{ob}-tree to make use of their advantages, concerning both the area of the region and the geographic feature inside the regions with more semantic for the users.

3 Semantic B^{ob}-Tree

Much of the research has been done in spatial obfuscation [1, 2, 10, 19], but all of these obfuscation techniques just deal with the geometry of the obfuscated regions. In other words, these obfuscation techniques concern only the area of the regions, but do not care about the geographic feature inside the obfuscated regions. Besides, each semantic obfuscation techniques has its disadvantage, technique in [3] does not concern the area; B^{ob}-tree has not enough semantic information attached for various expectation of different users. So, in this work, we propose a new semantic-aware obfuscation technique which concerns both the area of the regions and the geographic features inside the regions. This new technique not only ensures the same quality of service as others in [1] (because the obfuscated regions produced by these techniques have the same area), but also has higher user's location privacy.

3.1 Concepts

In our proposed technique, the regions are divided into three geographic features: sensitive, non-sensitive and unreachable. A place is sensitive when the user does not want to reveal to be in that area. A place is unreachable when the user because of various reasons, cannot enter in. Otherwise a place is non-sensitive. Like B^{ob}-tree, the requirement is that the obfuscated regions generated by this technique contain only reachable places, and satisfied the privacy of each user's expectation. Next is some concepts proposed in [3] which be used in our new technique.

Sensitive Level. Sensitive level defines the degree of sensitivity of a region for a user. It depends on the extent and nature of the objects located in the region as well as the privacy concerns of the user. It means that if a user is a doctor, a hospital where he/she works in not has a high sensitive level. But for another one, hospital has a high sensitive level because he/she wants to keep the health status to be secret. Sensitive level is in the range [0, 1]. Value 0 means that region is not sensitive or unreachable, we can publish the location of user is that region while value 1 means that region has the highest sensitivity and we have to hide that location.

Sensitive Threshold Value. The sensitive threshold value quantifies the maximum acceptable sensitivity of a location for the user. Its value ranges in [0, 1]. Value 1 means that the user does not care of location privacy, everyone can know his/her true position. Value is closer to 0 means higher degree of location privacy that user wants. A region r is location privacy preserving when its sensitive level is equal or smaller than the threshold value.

Feature Type. Any place belongs to a unique feature type. Users can specify the feature type that they consider sensitive, non sensitive and unreachable. A feature type is sensitive when it denotes a set of sensitive places. For example if hospital is a sensitive feature type, then Hospital A, an instance of Hospital, is a sensitive place. Instead a feature type is unreachable when it denotes a set of places which for various reasons, the user cannot enter in it. For example, the feature type Military Zone may be unreachable if the user is a common citizen. Otherwise, a place is non sensitive. The score of a feature type ft, score(ft), is used to specify how much sensitive ft is for the user. It is in the range [0, 1] and has the same meaning with sensitive level.

Privacy Profile. Users specify which feature types they consider sensitive and score of the sensitivity as well as the threshold value in a privacy profile. Every user has a particular privacy profile (or a group of users has one profile). We use this privacy profile to compute obfuscated regions that satisfy the privacy users expected.

Computation of Sensitivity Level. The sensitivity level (SL) of a region r, written by $SL_{Reg}(r)$ is defined in [3]. It is the sum of ratios of weighted sensitive area to the relevant area in the region. The weighted sensitive area is the surface in r occupied by sensitive feature weighted with respect to the sensitivity score of each feature type. The relevant area ($Area_{Rel}(r)$) of r is the portion of the region not occupied by unreachable feature. In the other words, the sensitivity level of a region r is defined by:

$$SL_{Reg}(r) = \Sigma\ Score(ft_{sens}) * (Area_{Fea}\ (r, ft_{sens}) / Area_{Rel}(r)), \text{ where: } Area_{Rel}(r) = Area(r) - \Sigma\ Area_{Fea}\ (r, ft_{UnReach}) \quad (1)$$

If r only contain unreachable features, we define $SL_{Reg}(r) = 0$.

For example, consider a region r which area is 350 (Area(r) = 350). The score of each feature type and its area in r are as figure 1.

We have: Area $(r, ft_1) = 50$, Area $(r, ft_2) = 100$, Area $(r, ft_3) = 200$, Area $(r, ft_4) = 0$. Apply the formula (1), the sensitive level of region r in figure 1 is:

$Area_{Rel}(r) = 100 + 200 + 0 = 300.$

$SL_{Reg}(r) = 0.9*100/300 + 0.5*200/300 = 0.633$

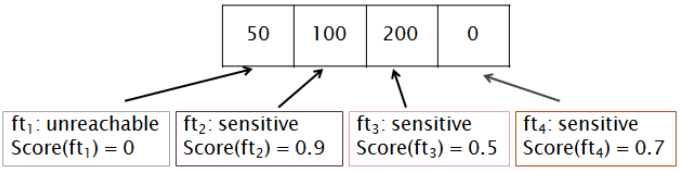


Fig. 1. Example of Computation of Sesitive level

Note that, the sensitive level of a region is less or equal to any sub region in it. For more details and proofs, please refer to [3]. It means that when we merge a region with another region, the sensitive level of the result region is always less or equal to the starting regions.

3.2 Index Structure

The structure of the Semantic B^{ob}-tree is based on B^{+}-tree which indexes one-dimensional values. Similar to B^{dual}-tree [5] and B^{ob}-tree [6, 7, 12], let o be a moving point with a reference timestamp $o.t_{ref}$, coordinates o[1], o[2],…,o[d], and velocities o.v[1], o.v[2],…,o.v[d], its dual is a 2d-dimensional vector as follows expression:

$o^{dual} = (o[1](T_{ref}),\ldots, o[d](T_{ref}), o.v[1],\ldots, o.v[d])$, where $o[i](T_{ref})$ is the i-th coordinate of o at time T_{ref} and is given by: $o[i](T_{ref}) = o[i] + o.v[i] * (T_{ref} - o.t_{ref})$ (2)

First, we apply the Hilbert curve to transform n-dimensional points to one-dimensional values. And then index these values into the structure of Semantic B^{ob}-tree like that of B^{+}-tree. However, each node of the tree has more information to assure semantic-aware privacy preserving. Beside area of reachable regions corresponding to the Hilbert range like B^{ob}-tree, its internal nodes (excluding the leaves) must contains the sensitive level of these regions.

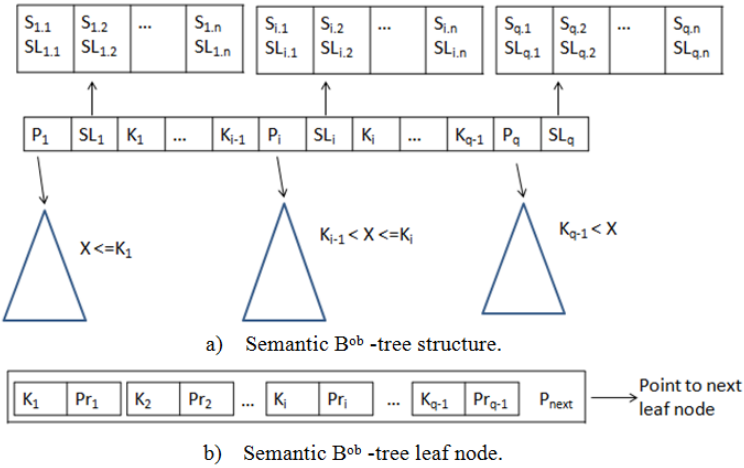


Fig. 2. Semantic B^{ob}-tree index structure

Figure 2 illustrates the structure of the Semantic B^{ob}-tree. Each internal node has the form $<P_1, SL_1, K_1, P_2, SL_2, K_2, \ldots, P_{q-1}, SL_{q-1}, K_{q-1}, P_q, SL_{q-1}>$ where P_i is the tree pointer, SL_i is pointer to an array contains S_i: the area of the reachable region associated with Hilbert interval $[K_{i-1}, K_i]$ and sensitive level (SL_i) of that region for each user (the number of element in this array is equal to number of users) and K_i is the search key value. Note that area of reachable regions S is also the relevant area concept in computation of sensitive level mentioned in equation (1).

Each leaf node has the form $<<K_1, Pr_1>, <K_2, Pr_2>, \ldots, <K_{q-1}, Pr_{q-1}>, P_{next}>$ where Pr_i is data pointer and P_{next} points to the next leaf node of the Semantic B^{ob}-tree.

The sensitive level of a region associated with each internal node (SL_i) will be calculated by applying the equation (1). The area of a region associated with each internal node (S_i) can be calculated by multiplying the total number of cells of each internal node with the area of the projection of each cell into coordinate space [6, 7, 12].

Figure 3 is a simple example of Semantic B^{ob}-tree structure with just one user.

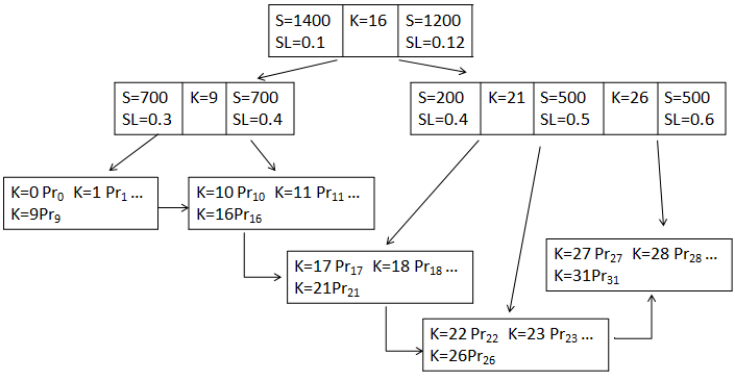


Fig. 3. An example of Semantic B^{ob}-tree with one user

The area of an obfuscated region associated with each internal node in a Semantic B^{ob}-tree is smaller and sensitive level is greater when traversing from the root to the leaf nodes, while the accuracy of user position increases and vice versa. Based on this property, the search process can stop at some internal nodes close to the root if the user wants the very high location privacy.

3.3 Search, Insert, Delete and Update Algorithms

Algorithm: Search
Input: a dual vector of a moving point o^{dual}, area of an obfuscated region S, sensitive thresh hold value θ.
Output: the region that its area equal to S and has sensitive level smaller than θ and contains a moving point with a dual vector o^{dual}.
Transform o^{dual} into Hilbert value h
while (n is not a leaf node) do
Search node n for an entry i such that $K_{i-1} < h <= K_i$

```
   if θ >= SL_i  then
        if S_i  = S then
           return the region corresponding to the Hilbert
interval [K_i-1, K_i].
        else if S > S_i  then
           return ExtendCell (K_i-1, K_i, S).
        else  // if S < S_i
          n ← n.P_i   // the i-th tree pointer in node n
    else
        return no solution.
search leaf node n for an entry (K_i, Pr_i) with h = K_i.
if found then retrieve the user's exact location.
else the search value h is not in database.
```

The algorithm ExtendCell (K_i, K_j, S) extends the region corresponding to the Hilbert interval [K_i, K_j] by adding more adjacent reachable cells until the area of the extended region equals to S. The sensitive level of the new region is equal or smaller than the old ones, thus it is still equal or smaller than the threshold value. This ensures that the obfuscated region generated by this technique has the same area as the geometry-based techniques and achieves better location privacy protection, because its concerns about the semantic of all objects in the region.

In this search algorithm, if the sensitive threshold value is small and the area S is big (e.g. user want the high degree of location privacy), the search process can stop at the internal node near the root node, we don't have to traverse to the leaf node to find the exact user's position as two phases techniques. Only in cases that users are willing to reveal their exact location, the search process must traverse to the leaf node.

The insert, update and delete algorithm of Semantic B^{ob}-tree are similar to those of B^{+}-tree. However, we have to recalculate S and SL of the region associated with each internal node. Due to the space limitation, we do not present the details here.

3.4 Simple Scenario

The map will be divided into cells. Beside the Hilbert values, each cell also contains sensitive information for each user. The users specify the feature types, which feature types they consider sensitive, unreachable, the sensitive score of each feature type and the threshold value in a privacy profile at the first time they register in the LBS system. They can update this profile later. Based on that information and users location, we generate and update Semantic B^{ob}- tree for everyone can request the services.

When the user wants to request services, he/she issues an authorization in form <id_{sp}, id_{user}, Δs> where id_{sp} is the identity of the service provider, id_{user} is the user's identity, and Δs is the area of the reachable regions. The result returns to the service provider, if any, is a reachable region with has the area of Δs, sensitive level of this region is equal or smaller than the threshold value and contains the user's position. If no solution is found because of cannot satisfy both area and sensitive threshold value, user may input another suitable area or change information in the privacy profile.

Besides, it can be interacted with the LBS middleware in [16] and queries processing techniques in [15] to operate completely.

4 Evaluation

4.1 Privacy Analysis

In [1], the authors have introduced the concept of relevance and accuracy degradation for location privacy measurement, as following:

$$\lambda = (A_i \cap A_f)^2 / (A_i \,.\, A_f) \tag{3}$$

In above formula, A_i is the area of location measurement returned by sensing technology based on cellular phones and A_f is obfuscated area that satisfied user's privacy. If the adversary may increase the value of λ from A_f , the privacy is decreased. In our context, assume that the natural degradation due to the intrinsic measurement error (the relevance associated with A_i) is small and A_i is included in A_f, so:

$$\lambda = A_i/A_f \tag{4}$$

Because the obfuscated area A_f of our technique just contains reachable regions and cannot reduce to any small area, the accuracy degradation λ of Semantic B^{ob}-tree is the same as equation (4).

In the geometry-based techniques, because the adversary can remove the unreachable region in A_f, we call the obfuscated area after the removing is A_{fg} ($A_{fg} <= A_f$), so:

$$\lambda_g = A_i/A_{fg} \tag{5}$$

Because $A_{fg} <= A_f$, from (4) and (5) we have $\lambda_g >= \lambda$, means that privacy of the Semantic B^{ob}-tree is higher than geometry-based techniques.

Similarly, in B^{ob}-tree, since the criteria of approachable and unapproachable parts division are completely based on geographic features, the adversary may reduce A_f to A_{fb} ($A_{fb} <= A_f$), by removing the unreachable regions because of other reasons except for geographic reasons (such as personal reasons). We have:

$$\lambda_b = A_i/A_{fb} \tag{6}$$

Since $A_{fb} <= A_f$, from (4) and (6) we have $\lambda_b >= \lambda$, means that the privacy of the Semantic B^{ob}-tree is also higher than B^{ob}-tree.

4.2 Performance

Intuitively, our new technique requires higher storage cost than B^{ob}-tree because it has to store sensitive information for each user in each internal node. Each internal node of Semantic B^{ob}-tree need more (16 bytes * number of users) to store area and sensitive level information (assume we use long and double data type for S and SL).

In the experiment, both B^{ob}-tree and Semantic B^{ob}-tree is all implemented in Java. The user's position, user's privacy profiles are randomly generated. Figure 4 shows the insert, search, update and delete cost (in milisecond) of B^{ob}-tree and Semantic B^{ob}-tree. We can see that the search time of two techniques is approximate. But the cost for inserting, updating and deleting items in Semantic B^{ob} -tree climbs steadily with increasing number of users.

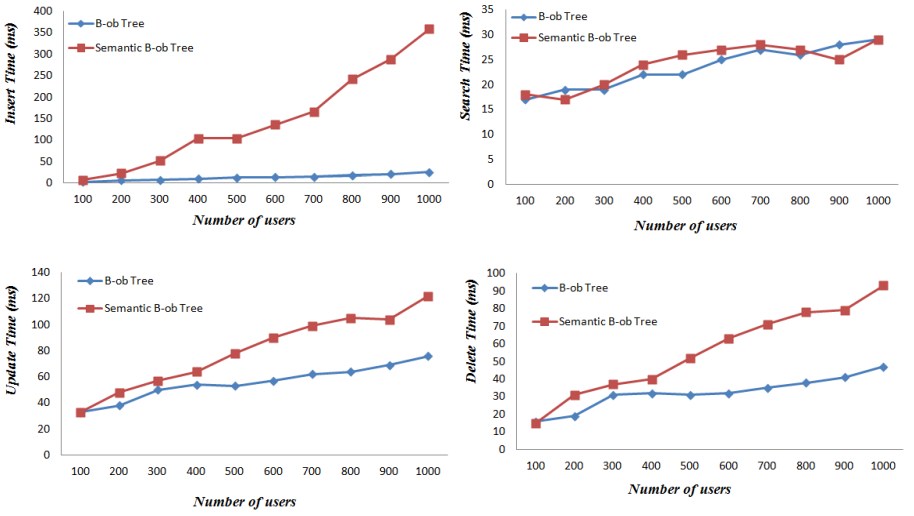


Fig. 4. Insert, search, update and delete time

5 Conclusion and Future Work

In this work, we have introduced the Semantic B^{ob}-tree, a new obfuscated technique for location privacy at database level. Theoretical analyses and discussions have shown that the newly proposed semantic-aware index structure can address the user location privacy more effectively than the B^{ob}-tree [6, 7, 12], which is the first index structure for obfuscation at database level. It is more concretely, more flexible and higher privacy.

In the future, we will intensively evaluate this technique using a variety of datasets to make a comparison with other techniques and optimize the index structure to increase the performance. Quality of services is also a big problem to consider next. Another research direction is to estimate the area and sensitive level of regions into one value for easily computation.

References

1. Ardagna, C.A., Cremonini, M., Vimercati, S.D.C., Samarati, P.: An Obfuscation-based Approach for Protecting Location Privacy. IEEE Transactions on Dependable and Secure Computing (2009)

2. Mohamed, F.M.: Privacy in Location-based Services: State-of-the-art and Research Directions. In: 8th International Conference on Mobile Data Management, Germany (2007)
3. Damiani, M.L., Bertino, E., Silvestri, C.: Protecting Location Privacy through Semantics-aware Obfuscation Techniques. In: IFIP Int. Federation for Information Processing (2008)
4. Duckham, M., Kulik, L.: A Formal Model of Obfuscation and Negotiation for Location Privacy. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 152–170. Springer, Heidelberg (2005)
5. Yiu, M.L., Tao, Y., Mamoulis, N.: The B^{dual}-Tree: Indexing Moving Objects by Space-Filling Curves in the Dual Space. VLDB Journal 17(3), 379–400 (2008)
6. To, Q.C., Dang, T.K., Küng, J.: B^{ob}-Tree: An Efficient B^+-Tree Based Index Structure for Geographic-Aware Obfuscation. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part I. LNCS, vol. 6591, pp. 109–118. Springer, Heidelberg (2011)
7. Dang, T.K., Thoai, N., To, Q.C., Phan, T.N.: A database-centric approach to privacy protection in location-based applications. In: RCICT, Lao PDR, pp. 65–71 (March 2011)
8. Gedik, B., Liu, L.: Protecting Location Privacy with Personalized k- Anonymity. IEEE Transactions on Mobile Computing 7(1), 1–18 (2008)
9. Kupper, A.: Location-based Services-Fundamentals and Operation (2005)
10. Truong, A.T., Truong, Q.C., Dang, T.K.: An Adaptive Grid-Based Approach to Location Privacy Preservation. In: Nguyen, N.T., Katarzyniak, R., Chen, S.-M. (eds.) Adv. in Intelligent Inform. and Database Systems. SCI, vol. 283, pp. 133–144. Springer, Heidelberg (2010)
11. Ardagna, C.A., Cremonini, M., Vimercati, S.D.C., Samarati, P.: Privacy-enhanced Location-based Access Control. In: Handbook of Database Security-Applications and Trends, pp. 531–552. Springer (2008)
12. To, Q.C., Dang, T.K., Küng, J.: A Hilbert-based Framework for Preserving Privacy in Location-based Services. Int. Journal of Intelligent Information and Database Systems (2013) ISSN 1751-5858
13. Phan, T.N., Dang, T.K.: A Novel Trajectory Privacy-Preserving Future Time Index Structure in Moving Object Databases. In: Nguyen, N.-T., Hoang, K., Jędrzejowicz, P. (eds.) ICCCI 2012, Part I. LNCS (LNAI), vol. 7653, pp. 124–134. Springer, Heidelberg (2012)
14. Verykios, V.S., Damiani, M.L., GkoulalasDivanis, A.: Privacy and Security in Spatiotemporal Data and Trajectories. In: Mobility, Data Mining and Privacy: Geographic Knowledge Discovery, ch. 8, pp. 213–240. Springer (2008)
15. Ngo, C.N., Dang, T.K.: On Efficient Processing of Complicated Cloaked Region for Location Privacy Aware Nearest-Neighbor Queries. In: Mustofa, K., Neunold, E., Tjoa, A M., Weippl, E., You, I. (eds.) ICT-EurAsia 2013. LNCS, vol. 7804, pp. 101–110. Springer, Heidelberg (2013)
16. Dang, T.K., Ngo, C.N., Phan, T.N., Ngo, N.N.M.: An Open Design Privacy-enhancing Platform Supporting Location-based Applications. In: ACM ICUIMC 2012. Springer, Kuala Lumpur (2012)
17. Truong, A.T., Dang, T.K., Küng, J.: On Guaranteeing k-Anonymity in Location Databases. In: Hameurlain, A., Liddle, S.W., Schewe, K.-D., Zhou, X. (eds.) DEXA 2011, Part I. LNCS, vol. 6860, pp. 280–287. Springer, Heidelberg (2011)
18. Dang, T.K., Truong, A.T.: Anonymizing but Deteriorating Location Databases. Int. Research Journal of Computer Science and Computer Engineering with Applications (POLIBITS) (46), 73–81 (2012) ISSN 1870-9044
19. Truong, Q.C., Truong, A.T., Dang, T.K.: The Memorizing Algorithm: Protecting User Privacy in Location-Based Services using Historical Services Information. Int. Journal of Mobile Computing and Multimedia Communications (a slected paper of MoMM 2009) 2(4), 65–86 (2010)

Practical Construction of Face-Based Authentication Systems with Template Protection Using Secure Sketch

Tran Tri Dang, Quynh Chi Truong, and Tran Khanh Dang

Faculty of Computer Science & Engineering, Ho Chi Minh City University of Technology, VNUHCM, Vietnam
{tridang,tqchi,khanh}@cse.hcmut.edu.vn

Abstract. Modern mobile devices (e.g. laptops, mobile phones, etc.) equipped with input sensors open a convenient way to do authentication by using biometrics. However, if these devices are lost or stolen, the owners will confront a highly impacted threat: their stored biometric templates, either in raw or transformed forms, can be extracted and used illegally by others. In this paper, we propose some concrete constructions of face-based authentication systems in which the stored templates are protected by applying a cryptographic technique called secure sketch. We also suggest a simple fusion method for combining these authentication techniques to improve the overall accuracy. Finally, we evaluate accuracy rates among these constructions and the fusion method with some existing datasets.

Keywords: Secure sketch, Face authentication, Biometric template protection.

1 Introduction

Current mobile devices (laptops, mobile phones, etc.) are used not only for simple tasks like communicating or web browsing, but they can also be used to do more complex tasks such as learning and working. As a result, some sensitive information is stored on mobile devices for such tasks. To ensure the confidentiality of the personal information, usually an authentication process is implemented. Besides password-based authentication, modern devices equipped with input sensors open a new way of doing authentication: biometric. Using biometric for user authentication actually has some advantages [1]. However, while passwords can be easily protected by storing only their one-way hash values, it is not easy to do so with biometric data. The problem lies in the noisy nature of biometric data, i.e. the biometric templates captured from the same person in different times will certainly be different. Hence, if we apply the one-way hash function to biometric data, we will unable to compare the distance between the stored data and the authentication data.

In this paper, we propose one construction that offers face-based authentication and provides protection for stored biometric templates at the same time. We follow the concept of "secure sketch" proposed by Dodis et al. [2]. One property of secure sketch is that it allows reconstructing the original biometric template exactly when

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 121–130, 2013.

provided another template that is closed enough to the first one. Because of that property, we can protect the stored templates by using a one-way hash function on them.

The remains of this paper are structured as follow: in section 2, we review some related works; in section 3, we present our construction technique of secure sketch on 3 different face recognition algorithms; in section 4, our experiment results are reported and base on them we introduce a simple fusion method to improve the performance of our system; finally, in section 5, we conclude the paper with findings and directions for future researches.

2 Related Works

Using biometric for authentication is not new [3]. One perspective of this problem is how to reliably and efficiently recognize and verify the biometric features of people. This is an interesting topic for pattern recognition researchers. Another perspective of this problem is how to protect the stored templates and this is the focus of security researchers as well as this paper.

There are many approaches to the problem of template protection for biometric data. One of which is the "secure sketch" proposed by Dodis et al. [2]. In its simplest form, the working of the secure sketch is described in Fig.1.

Fig. 1. The working of the secure sketch

There are 2 components of the secure sketch: the sketch (*SS*) and the recover (*Rec*). Given a secret template *w*, the *SS* component generates public information *s* and discards *w*. When given another template *w'* that is closed enough to *w*, the *Rec* component can recover the *w* exactly with the help of *s*. There are 3 properties of the secure sketch as described in [2]:

1. *s* is a binary string {0, 1}*.
2. *w* can be recovered if and only if $|w - w'| \leq \delta$ (δ is a predefined threshold).
3. The information about *s* does not disclose much about *w*.

In our construction, only property 2 is guaranteed. Fortunately, we can easily transform our sketch presentation into binary string to make it compatible with property 1. And although we do not prove property 3 in this paper, we do give a method to measure the reduction of the search space used to brute-force attack this system using public information s. For these reasons, we still call our construction secure sketch.

Our construction is implemented with biometric templates extracted from 3 face recognition methods: the Eigenfaces method proposed by Turk and Pentland [4], the 2DPCA method proposed by Yang et al. [5], and the Local Binary Patterns Histograms (LBPH) proposed by Ahonen et al. [6]. We select more than one face

recognition method to experiment how generic our construction is when applied to different template formats. Another reason is we want to combine the results of these individual authentications to improve the overall resulted performance.

To recover *w* exactly when given another *w'* closed to it, some error correction techniques are needed. In fact, the public information *s* is used to correct *w'*, but it should not disclose too much about *w*. We follow the idea presented in [7] paper to design this error correction technique. Our technique can be applied on discrete domains and it gives reasonable results when experimenting with the Eigenfaces, 2DPCA, and LBPH face recognition methods.

Individual biometric recognition systems can be fused together to improve the recognition performance. The fusion can be implemented at feature extraction level, score level, or decision level [8]. In this paper, based on the specific results obtained from the experiments with individual features (i.e. Eigenfaces, 2DPCA, and LBPH), we propose a simple fusion technique at the decision level, and in fact, it improves the overall performance significantly.

3 Construction Methods

3.1 Processing Stages

The stages of our construction are summarized in Fig. 2. Firstly, the face feature is extracted. The formats of the extracted features depend on the face recognition methods used. Secondly, a quantization process is applied to the features' values in continuous domain to convert them into values in discrete domain. This stage is needed because discrete values allow exact recovery more easily than continuous values do. The quantized values play the role of *w* and *w'* as described in previous section. The sketch generation stage produces *s* given *w*. And finally, the feature recovery stage tries to recover the original *w* given another input *w'* and *s*. To validate whether the recovered feature matches with the original feature *w*, a one-way hash function can be applied to *w* and the result is stored. Then, the same hash function will be applied to the recovered feature and its result is compared with the stored value.

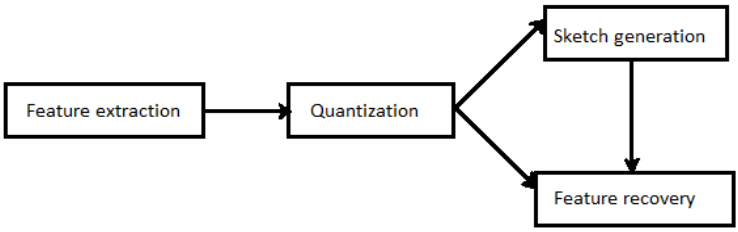


Fig. 2. The stages of our construction

3.2 Feature Extraction

Eigenfaces

Given a set of training face images, the Eigenfaces method finds a subspace that best represents them. This subspace's coordinates are called eigenfaces because they are

eigenvectors of the covariance matrix of the original face images and they have the same size as the face images'. The detail of the calculating these eigenfaces was reported in [4].

Once the eigenfaces are calculated, each image can be projected onto its space. If the number of eigenfaces is N, then each image in this space is presented by an N-dimensional vector.

2DPCA

The 2DPCA [5] works similarly to the Eigenfaces. However, while the Eigenfaces treats each image as a vector, the 2DPCA treats them as matrixes. Then, the 2DPCA tries to find some projection unit vectors X_i such that the best results are obtained when the face matrixes are projected on them.

In 2DPCA, the projection of a face matrix M on a vector X_i resulted in a transformed vector Y_i that has the number of elements equals to the rows of M. If a face matrix has R rows, and the number of projection vectors is P, then the transformed face matrix has a size of RP. In other words, each face image in the 2DPCA method is presented by an N-dimensional vector, in this case $N = RP$.

Local Binary Patterns Histogram

Local Binary Patterns (LBP), which was first introduced by Ojala et al. [9], is used for texture description of images. The LBP operator summarizes the local texture in an image by comparing each pixel with its neighbors. Later, Ahonen et al. proposed LBPH method for face recognition based on the LBP operator [6].

In this method, at first, a face image is converted to LBP image. Each pixel value is computed by its neighbor values. If the center pixel is greater or equal its neighbor value, then denote it with 1 and 0 otherwise. The surrounding pixels yield a binary number for a center pixel (Fig. 3). After that, the image is divided into small areas and histograms are calculated for each area. The feature vector is obtained by concatenating the local histograms. In this case, if an image is divided into A areas, then it is presented by an N-dimensional vector, in this case $N = 256A$ (256 grayscale values).

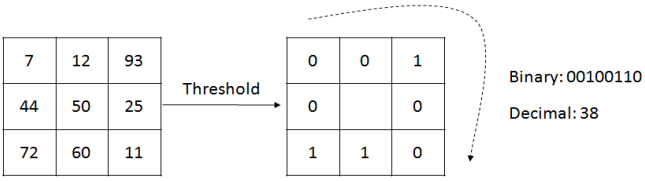


Fig. 3. LBP operator

3.3 Quantization

The purpose of the quantization stage is to convert feature values from continuous domain to discrete domain. At a first glance, this stage may reduces the security of the authentication process significantly by reducing the continuous search space with infinite elements to a discrete search space with finite elements. However, an

informed attacker will understand that biometric authentication is not an exact-matching process, and therefore no need to try every possible values, but only values separated by a threshold. In other words, only finite values are needed to brute-force attacks the continuous template values. Furthermore, we can control the size of the quantized domain by changing the range of the continuous values that mapped to the same quantized value. For these reasons, this stage actually does not affect the security of the authentication system.

Our quantization process works as follow: after normalization, the value of each element of feature vectors is a floating point number in [0, 1]. Then, the quantization process will transform this value to an integer in [0, N], with $N > 0$. Let x be the value before quantization, and x' be the value after quantization, then the quantization formula can be written as (1). *Round* function returns a nearest integer to a parameter.

$$x' = round(xN) \tag{1}$$

3.4 Sketch Generation

The sketch generation stage produces public information s that can be used later to recover quantized template w. Our construction, based on the idea presented in [7] paper, is described below.

The domain of w is [0, N]. We create a codebook where the codewords spread along the range [0, N] and the distance between any pair of neighbor codeword is the same. In particular, the distance between the codeword c_i and c_{i+1} is 2δ where δ is a positive integer. Then, for any value of w in the range $[c_i - \delta, c_i + \delta]$, the mapping function M returns the nearest codeword of w, or $M(w) = c_i$. The sketch generation use the mapping function to return the difference between a value w and its nearest codeword, or

$$SS(w) = w - M(w) \tag{2}$$

The values of the sketch generation $SS(w)$ is in the range $[-\delta, \delta]$ irrespectively of the particular value of w. So, given $SS(w)$, an attacker only knows that the correct value w is in the form $SS(w) + M(w)$. To brute-force attack the system, the attacker needs to try all possible values of $M(w)$, or every codeword. The larger δ is, the smaller the codeword space is. Note that the codeword space is always smaller than the quantized space [0, N]. When $\delta = 1$, the codeword space is three times smaller than the quantized space, and when $\delta = 2$, this number is five times.

3.5 Feature Recovery

Given the authentication input w', the feature recovery stage uses s and w' to reproduce w if the difference between w and w' is smaller than or equal to δ. Call the recovered value w'', it is calculated as

$$w'' = M(w' - SS(w)) + SS(w) \tag{3}$$

To prove the correct w is reproduced when $|w - w'| \leq \delta$, replace $SS(w)$ by the right-hand-side in (2), we have

$$w'' = M(w' - w + M(w)) + w - M(w) \quad (4)$$

If $|w - w'| \leq \delta$, then $w' - w + M(w)$ is in $[M(w) - \delta, M(w) + \delta]$

According to the codebook construction, applying the mapping function on any value in this range will return its nearest codeword, which is also $M(w)$. Substituting the function $M(w' - w + M(w))$ as $M(w)$ in formula (4), we have $w'' = w$.

3.6 Security Analysis

In this section, we consider the security of our proposed construction against the most basic attack: brute-force. Each feature template w can be considered as a vector of N elements. To get the correct w, every element in the vector must be corrected successfully. In previous section, we demonstrated that every codeword must be tried to brute-force attack an element. In current construction, we use the same quantization and codebook for every element regardless of its value distribution. So, assume there are S codewords in a codebook, and then in average, an attacker must try $\frac{S^N}{2}$ cases to get the correct w using a brute-force attack. Of course, the attacker may use the distribution information of the values to reduce the number of searches needed, but it is out of the scope of this paper.

Here are some specific numbers regarding the security of our experiments

- The quantized range is [0, 1000]
- Eigenfaces:
  - We tested with $N = 10, 12, 14, 16, 18$
  - Codeword space range from $S = 20$ (with $\delta = 25$) to 200 (with $\delta = 2$)
  - The minimum and maximum security offered are 20^{10} and 200^{18}
- 2DPCA
  - Image height is 200 and the number of projection axes chosen is 1, 2, 3, 4, 5, and 6. So, we have $N = 200, 400, 600, 800, 1000$, and 1200 respectively
  - Codeword space range from $S = 2$ (with $\delta = 250$) to 10 (with $\delta = 50$)
  - The minimum and maximum security offered are 2^{200} and 10^{1200}
- LBPH
  - We divide the images into a 5 x 5 grid. So, we have $N = 5 \times 5 \times 256 = 6400$
  - Codeword space range from $S = 5$ (with $\delta = 100$) to 10 (with $\delta = 50$)
  - The minimum and maximum security offered are 5^{6400} and 10^{6400}

4 Experiments

Our proposed constructions are then tested with the Faces94 database [10]. The experiments measures true accept rates (the percentage of times a system correctly accepts a true claim of identity) and true reject rates (the percentage of times a system

correctly rejects a false claim of identity) of different recognition algorithms and with different codeword space. The purpose of the experiments is to verify an ability to apply these constructions in real applications with reasonable threshold. We choose images of 43 people, randomly in the Faces94 database. We have 2 sets of images per a person, one for creating feature vectors and one for recovering feature vectors. For each algorithm, we will conduct 43 x 43 tests, in which 43 of them (testing a person with himself/herself) should recover and 1806 (testing a person with others) should not.

4.1 Individual Tests

Eigenfaces

Firstly, we choose mages of 37 people (2 images from each person, 17 are female) are selected randomly from the Faces94 database to create the eigenfaces. They need not to be the same as 43 people in the training set. Next, we use the first set of 43 people to create feature vectors. Then, the other set is used to recover the original feature vectors. For every pair of image, x and y, the scheme try to recover x from y. So, if x and y are the same, the system should recover correctly and if x is different from y, the system should not be able to recover x. The eigenfaces numbers chosen are 10, 12, 14, 16 and 18. And the codeword are chosen with space equally and δ range from 2 to 25. Our true reject rate is always 100%, so we only show the performance in term of true accept rate. The true accept rate is depicted in Fig. 4.

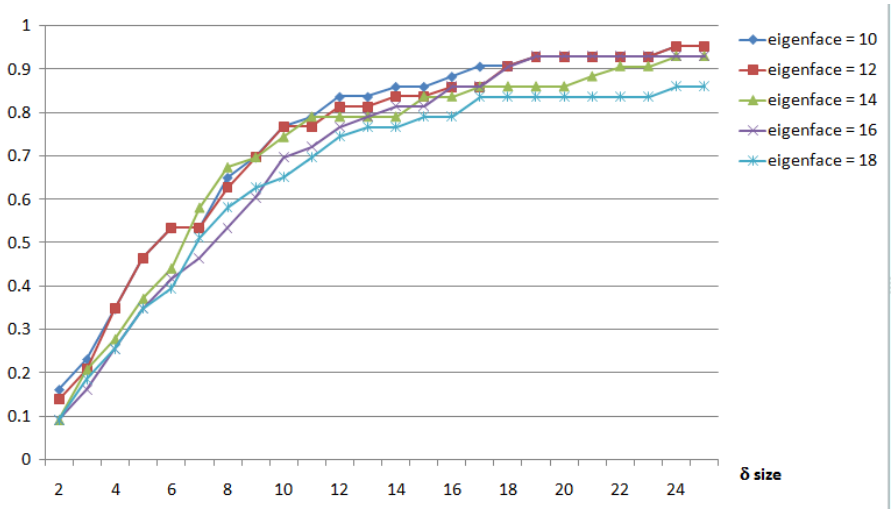


Fig. 4. True accept rate for Eigenfaces algorithm

2DPCA

The settings to measure the true accept rate and true reject rate of this experiment is similar to the settings in Eigenfaces. The number of projection axis chosen is just only 2 to 6, because just the image height is 200 pixels, a large enough dimension value. And the codeword are chosen with space equally and δ range from 50 to 250. As in

the experiment with Eigenfaces algorithm, our true reject rate is always 100%, so we only show the performance in term of true accept rate. The true accept rate of this system is depicted in Fig. 5.

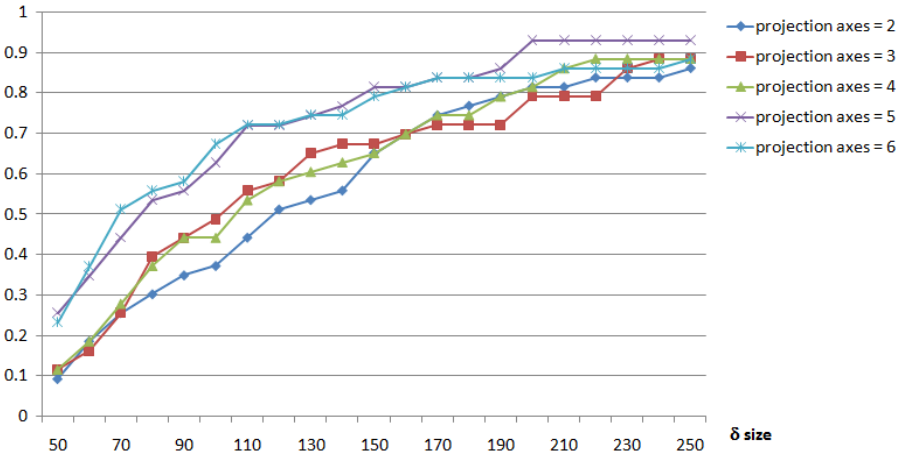


Fig. 5. True accept rate for 2DPCA algorithm

Local Binary Patterns Histograms

For the local histogram algorithm, we divide the face images into a 5 x 5 grid, and for each cell of the grid, there are 256 values for 256 grayscale levels. Hence, the dimension of the feature vector in this case is 5x5x256, which is 6400. The codeword are chosen with space equally and δ range from 50 to 100. The reason we stop at 100 is beyond this value, the true reject rate decrease significantly. Unlike the 2 previous algorithms, the local histogram returns true reject rate less than 100% when the size of δ is more than some threshold. The true accept rate and true reject rate for this algorithm is depicted in Fig. 6

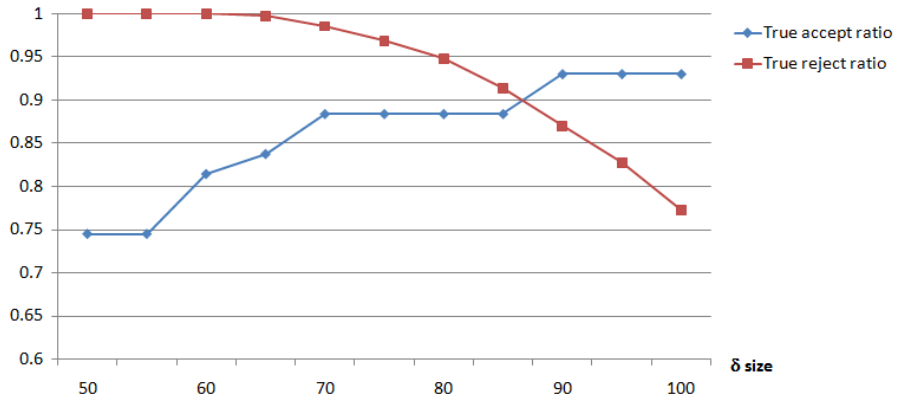


Fig. 6. True accept rate and true reject rate for LBPH algorithm

4.2 Fusion Tests

The working of our face-based secure sketch for Eigenfaces and 2DPCA algorithms both achieve the true reject rate of 100%, but none of them achieve the 100% true accept rate with different experiment parameters. To further experiment if the fusion of these results could improve the overall performance of our construction, we implement a simple fusion method of these 2 algorithms at the decision making level. In this case, we want to improve the true accept rate, so our fusion is the Boolean function OR that will return true when either the Eigenfaces or 2DPCA result matches. Because the Eigenfaces algorithm produces best result when the eigenfaces number is 12, and because the 2DPCA algorithm produces best result when the projection axes are 5, we use these setting in the fusion construction. The δ values for eigenfaces are chosen at 5, 10, 15, 20, and 25; the δ values for 2DPCA are chosen at 50, 75, 100, 125, 150, 175, 200, 225, and 250. The true reject rate of our fusion also return 100%, but there is a significant improvement in the true accept rate of the construction. In fact, when the δ value of the 2DPCA reach 100, the fusion always return 100% true accept rate when selecting the δ value for the Eigenfaces algorithm at 5, 10, 15, 20 and 25.

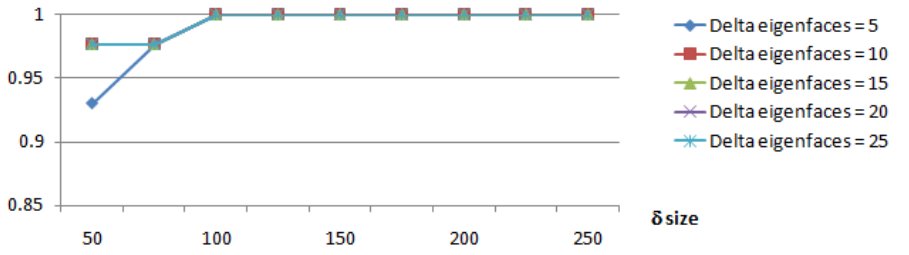


Fig. 7. True accept rate for Fusion test

5 Conclusions and Future Works

In this paper, we present a practical construction of face-based authentication technique with template protection using secure sketch. Although not exactly as the secure sketch proposed in the [2] paper, we demonstrate some security measure of our method in which brute-force is the only attacking technique used. Experiment results show the potential of our construction for using in security application, which the true reject rate is always 100%. The true accept rate of our construction is also increased when a simple fusion technique is applied.

However, more theoretical works is needed to prove, especially the security bound when attackers know about the distribution of the extracted feature values. Furthermore, the construction is also need to be tested on more complex human face database to see how it works. Not only improvement on individual feature authentication, there is a need to improve the fusion method. The fusion now is just a simple Boolean function at the decision level. When the feature is recovered correctly, it can be used to

calculate the distance between the original feature and the feature used for authentication. Using this distance in the fusion may give more choices in designing the final result. And finally, as the development of modern devices, more sensors input are equipped to capture other features, therefore fusion between different biometric features is also a possible way to enhance the system.

Acknowledgements. The authors would like to give special thanks to POSCO, South Korea, for their financial support.

References

1. O'Gorman, L.: Comparing Passwords, Tokens, and Biometrics for User Authentication. Proceedings of the IEEE 91(12), 2021–2040 (2003)
2. Dodis, Y., Ostrovsky, R., Reyzin, L., Smith, A.: Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data. SIAM J. Comp. 38(1), 97–139 (2008)
3. Wayman, J., Jain, A., Maltoni, D., Maio, D. (eds.): Biometric Systems: Technology, Design and Performance Evaluation. Springer, London (2005)
4. Turk, M., Pentland, A.: Eigenfaces for Recognition. J. Cognitive Neuroscience 3(1), 71–86 (1991)
5. Yang, J., Zhang, D., Frangi, A.F., Yang, J.-Y.: Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition. IEEE Trans. Pattern Anal. Mach. Intell. 26(1), 131–137 (2004)
6. Ahonen, T., Hadid, A., Pietikäinen, M.: Face Recognition with Local Binary Patterns. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
7. Juels, A., Wattenberg, M.: A Fuzzy Commitment Scheme. In: 6th ACM Conference on Computer and Communications Security, pp. 28–36. ACM, New York (1999)
8. Ross, A., Jain, A.: Information Fusion in Biometrics. Pattern Recogn. Lett. 24(13), 2115–2125 (2003)
9. Ojala, T., Pietikäinen, M., Harwood, D.: A Comparative Study of Texture Measures with Classification based on Feature Distributions. Pattern Recogn. 29, 51–59 (1996)
10. Libor Spacek's Faces94 database, `http://cswww.essex.ac.uk/mv/allfaces/faces94.html`

CAPTCHA Suitable for Smartphones

Yusuke Tsuruta, Mayumi Takaya, and Akihiro Yamamura

Akita University, Department of Computer Science and Engineering,
1-1, Tegata Gakuen-machi, Akita, 010-8502 Japan
tsuruta2013@gmail.com, {msato,yamamura}@ie.akita-u.ac.jp

Abstract. We propose a CAPTCHA that tells data made by a computer program from one-stroke sketch data given by a human being using embodied knowledge. Utilizing touchscreens of smartphones, we realize this approach and resolve a conceivable inconvenience caused by the existing CAPTCHAs when using smartphones due to the limited display size of smartphones. We implement the proposed technique and analyze its validity, usefulness and security.

Keywords: CAPTCHA, Smartphones, Touchscreens, Embodied knowledge.

1 Introduction

A CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is one of the reverse Turing tests ([7]) that distinguish an access from a computer program such as a crawler from an access from human beings using the difference between humans' shape recognition ability and the machine recognitions ([2,3]). The computer program might access to network services to acquire a large amount of accounts of the web mail service aiming at an illegal network utilization such as sending spam mails or carrying out APT attacks. A CAPTCHA can be applied to a web security technique preventing such illegal accesses to network service. When facing the existing CAPTCHA, a human recognizes a word, maybe a nonsense sequence of characters, in the image on the display and is required to respond by typing the word through the keyboard (Fig. 1). The character image is distorted in some fashion, and computer programs cannot recognize easily the characters. For example, an OCR program cannot recognize the character, whereas human beings can do it without difficulty. CAPTCHAs have been analyzed and several methods based on different principles have been proposed ([4]).

Smartphones play an important role in the information-communication society nowadays, and the development of cloud computing promotes the spread of smartphones and has influenced the use of the Internet. Actually, the accesses from smartphones to internet services rapidly are increasing, and actually many web sites have begun to correspond to smartphone users. There are many differences between smartphones and the past computer models from the point of view of human-computer interface. For smartphones, data is inputted through

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 131–140, 2013.

the touchscreen by hands, and the display of a smartphone is comparatively small. A CAPTCHA login is requested when a user is accessing to internet services through a smartphone exactly same as it is requested to accesses from the desktop PCs. When we use smartphones and face a CAPTCHA, both the challenge image and the virtual keyboard are displayed, and we have to type in the word displayed in the image. However, the virtual keyboard occupies almost half of the display and so the CAPTCHA image must be small. To solve this problem, a new image based CAPTCHA is proposed in [5]. In this paper, we propose yet another CAPTCHA suitable for smartphones using embodied knowledge of human beings. Our approach is different form [5].

Embodied knowledge is the control of the muscle acquired by practice and the motor learning of the brain such as the skill remembered in childhood. Realizing embodied knowledge by a computer is one of the challenging problems in artificial intelligence research. The proposed technique is to decide whether or not the response to the challenge is created by human beings or computer programs by checking the existence of embodied knowledge. One-stroke sketch is taken up as an ingredient in the embodied knowledge of our proposal. Human-computer interaction through a touchscreen that is one of the features of smartphones is suitable for faithfully acquiring one-stroke sketch data. One-stroke sketch input data with humans' finger is characterized as a continuous locus resulted by the human hand's physicality and realized as a series of coordinates on the display. The entire character image is not drawn at the same time but it is drawn continuously on the curve along the shape of the character little by little following the tracks of the tip of a finger according to the operation of the arm and the hand.

It is also necessary to give data the continuous order at the pixel level so that the computer program may compose legitimate input data. Therefore, the proposing CAPTCHA is based on not only hardness of image recognition but also the embodied knowledge which is an important theme in artificial intelligence.

Fig. 1. A typical CAPTCHA

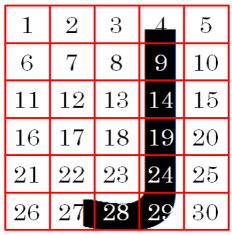


Fig. 2. Example("J")

2 The Proposed Technique

2.1 Basic Idea

As a standard authentication protocol, a CAPTCHA server sends a challenge and the user has to respond to it in a correct way. In the case of the proposed CAPTCHA, the server sends a challenge image that includes a character (or a symbol), and then the user is requested to trace the character by a finger tip and sends back the data representing a one-stroke sketch as a response to the challenge. The smartphone interprets the date inputted as an ordered series of coordinates in which the order is given as the time series. The server receives the ordered series of coordinates and check whether or not it is acceptable as a data obtained by a human using implicitly embodied knowledge. If the data received is determined as an output of a computer program then the access is rejected.

The touchscreen is delimited and the grid is composed. In tracing the character included in the image on the display by the finger, coordinates of the points in the display touched by the finger are acquired by the drug operation of the touchscreen. The server determines whether or not the series of coordinates is acceptable by checking the locus is correct and the data input is continuous in addition to the correctness of the the starting point and the terminal point.

For instance, suppose the image of the character "J" is displayed as a challenge in Fig. 2. The correct response is obtained by dragging the finger along the shape of "J" on a touchscreen of a smartphone. Coordinates of the input data, that is, the series of coordinates of the locus are checked if the first coordinate is included in the small area 4, and if the following coordinates are in the small area 9 and so on. If the series of ordered coordinates is nearly in the order of the small areas 4, 9, 14, 19, 24, 29, 28, and 27, it is accepted (Fig. 2). If the coordinates is in the order of the small areas 2, 8, 14, 19, and 20, then the data is rejected. We shall explain the proposed technique in detail in the expanded version of the paper.

2.2 Security

The main objective of CAPTCHAs is to prevent computer programs from accessing to network services for evil purposes. Therefore, the attacker is a computer program disguising as a human being and trying to obtain a legitimate authority to access. Then the security of a CAPTCHA technology is evaluated by the intractability for computer programs to obtain the access permit ([1,6]). We analyze conceivable attacks against the proposal CAPTCHA.

Suppose that the image displayed as a challenge is monochrome, and the character is drawn in the white ground in black. In this case, the coordinate data of the area where the character is drawn can be acquired accurately by examining RGB of the challenge image. Then a computer program should enumerate a series of coordinates at which black RGB is appointed and give order to these coordinates according to the correct writing, that is, following one-stroke sketch. For example, if a human write "J" then the input data trace the small areas 4,

9, 14, 19, 24, 29, 28, and 27 like in Fig. 2. The number of the coordinates should be nearly same as the standard input by human beings to disguise. In general, a computer program has no information on one-stroke sketch, which is considered as an embodied knowledge of human beings. Each human being has learned such an embodied knowledge from their childhood. A computer program should pick one position from the area of a coordinates with black RGB as the starting point and also as the terminal and then compose a series of coordinates with black RGB that connects the starting and terminal points in a correct order. It is impossible to execute this task if there is no information on the stroke order. If many responses are permitted to the same challenge image, the brute force attack becomes possible in principle. However, the brute force attack can be avoided by permitting only one response to each challenge. Moreover, it is realistic to put the limitation on the challenge frequency.

Now assume that an attack program has the database concerning characters and the correct order of writing. If a series of coordinates can be correctly obtained, then information on the correct order of writing might be able to be obtained from the database. We note that it costs a lot to make such a database for attack against the CAPTCHA and so this already has some deterrent effect. In addition, the challenge is not necessarily based on a character or a symbol. An arbitrary curve can be used for a challenge instead of a character and then making the database is impossible in principle. We shall discuss this issues in the future work. We may execute transformations on the shapes and colors of the character to perplex computer programs. If the transformation processing is a continuous transformation, this occurs no trouble for human beings and so such a transformation is allowed. It seems difficult for computer programs to respond correctly (Fig. 5). If the challenge is a (not necessarily monochrome) color image, the attacker's program has to carry out an edge detection and specify the character. Using the existing CAPTCHA techniques such as adding the distortion to the character, the attacker's program has the difficulty to detect the character. Moreover, not only adding the distortion transformation but also camouflaging the background with the dazzle paint makes the attacker's program hard to detect the character. Therefore, the security of the proposed CAPTCHA is at least the existing CAPTCHAs because their techniques can be employed to our CAPTCHA as well. In addition, the method requiring the user to input more than one stroke traces is effective to improve the security level. The security level can be adjusted according to the system requirement. To understand the security of the proposed CAPTCHA well, we should examine human embodied knowledge from the standpoint of the cognitive psychology.

Fig. 3. Separator Image

Fig. 4. Deformed Image

Fig. 5. Distorted Image

2.3 Comparison with the Existing CAPTCHA

We discuss the usefulness of the proposed CAPTCHA comparing with the existing techniques provided that a user is accessing using a smartphone. Note that the screen size of smartphones is about 3.5-5 inch. When using smartphones, both a CAPTCHA image and a virtual keyboard are displayed (Fig. 6) and the size of the CAPTCH image is almost half of the screen. It is very inconvenient for most of the users to respond to a CAPTCHA challenge due to this limited size image and the virtual keyboard. One has to type more than once to input a character when using a virtual keyboard. For example, when typing "c" in the lowercase letter, one has to press the button for "c" three times (Fig. 7). When typing "C' ' in the uppercase letter, one needs more operations to change the "lowercase mode" to the "uppercase mode". Therefore, the total number of operations becomes enormous if the words are arbitrarily generated using lowercase letters, uppercase letters and figures. Moreover, a wrong character may be inputted by an unintentional typing mistake. For this reason, some existing CAPTCHA use only figures $(0, 1, 2, \ldots, 9)$ without using alphabets to improve user's convenience. Note that if uppercase and lowercase letters are allowed in addition to the figures, $62 (= 10 + 52)$ characters can be used. This results in the deterioration of the security; if the challenge is a word consisting of n letters, there are only 10^n cases compared with 62^n cases.

When using the proposed CAPTCHA, the entire display is used for showing the challenge image, and the input is comparatively easy (Fig. 8); no additional operations such as changing modes are required.

Fig. 6. Existing(num)

Fig. 7. Existing(char)

Fig. 8. Proposed Method

3 Line Trace Attack

It may seem possible to use the line trace program, which is often used in a robot, to trace the black coordinate area of the challenge image for attacking the CAPTCHA. For this attack, the line trace program has to trace on the black area from the starting point of the character to the terminal point in order to compose the response data. The attack using a line trace program seems the most plausible attack as of now. It is necessary for a line trace program to find the starting point to begin the tracing, however, it seems intractable to find the starting

point because the line trace program checks the local area and determines the next action and the starting point is usually given as the input to the program by a human being. A human being looks at the image, comprehends the character and finds the starting point using the embodied knowledge. On the other hand, choosing a starting point is intractable for a line trace program. If a human takes part in the attack, the attacker consists of not only a program but a human, and so this approach is excluded as an attack against the proposed CAPTCHA. For an objective of a CAPTCHA is to prevent programs from accessing without human beings assistance. Even if the starting point is obtained in some ways without human assistance, our approach allows challenges such as separated images (Fig. 3) or deformed images (Fig. 4) to perplex the line trace program, which give no trouble to human beings as we see in the subsequent section. Therefore, an attack using line trace programs seems intractable.

3.1 Experiment of Attack Using Line Trace Program

In the following experiments of attacking against the proposed CAPTCHA, a line trace program tries to make an acceptable response to the challenge images (Fig. 3, 5, 9, 10, 11, 12, 13, 14). Each experiment is executed provided the starting point is given to the program beforehand by a human. We use a simulation line trace program [8] in this experiment.

Fig. 9. Test Image 1

Fig. 10. Test Image 2

Fig. 11. Test Image 3

Fig. 12. Test Image 4

Fig. 13. Test Image 5

Fig. 14. Test Image 6

The line trace program succeeded in making an acceptable response only to the challenge image in the image 4 (Fig. 12), and it failed against the other images (see Table 1). By these experiments, we conclude that countermeasures leading the line trace program to a dead end or putting the pause in the character shape are considerably effective whereas these do not cause any troubles to human beings. The line trace program also fails to trace when the angle formed in the character shape is too big. As we have already mentioned that the line trace

program is given the starting point as an input by human beings. However the actual attack must be carried out without human beings' assistance. Therefore, a simple attack using a line trace program does not seem a serious threat against the proposed CAPTCHA. We shall report the detail of the experiments and discuss more about the results in the expanded version of this paper.

Table 1. Correspondence Table

	Image1	Image2	Image3	Image4
Trace Success	-	-	-	√
Trace Failure	√	√	√	-
	Image5	Image6	Fig. 3	Fig. 5
Trace Success	-	-	-	-
Trace Failure	√	√	√	√

4 Validity of the Proposed CAPTCHA

We examine the validity of the proposed CAPTCHA by experiments; 22 subjects (humans) are asked to respond to several challenge images that represent the symbol "α".

4.1 Experiments

We use a handheld computer (Android 3.1 and processor NVIDIA Tegra 2 mobile processor) equipped with 9.4 type WXGA liquid crystal with the internal organs display touchscreen of the ITO Grid method mirror electrostatic capacity method as the user machine. The platform of the server is constructed on Windows 7 Professional 64bit, 2048MB memory, and Intel Core i3, and the authentication program is written by using c/c++ compiler MinGW. The size of the challenge images is 1200×700 pixel. The response is accepted if the locus is passing in a correct order.

The purposes of each experiment are summarized as follows (see Table 2). In the experiment 1 and 2, the small zone is set 35×35 pixels and 70×70 pixels, respectively, and we investigate the differences between these two cases. In the experiment 3, the small zone is set 35×35 pixels and we specify the entry speed and the input position and investigate the difference between these cases. In the experiment 4, we investigate the effect caused by the change of characters. In the experiment 5, we investigate the case that the response is accepted only when all set coordinates are passed. In the experiment 6, we investigate the tolerance for human non-intentional errors.

Experiment 1. The instruction "Please trace on the character shape by one stroke" is displayed and the challenge image Fig.9 is displayed. The small zone on the grid is 70×70 pixel(Fig.15).

Experiment 2. The instruction "Please trace on the character shape by one stroke" is displayed and the challenge image Fig.9 is displayed. The small zone on the grid is 35×35 pixel (Fig.16).

Experiment 3. The instruction "Please trace on the character shape by one stroke within 5 seconds" is displayed and the challenge image Fig.9 is displayed. The small zone on the grid is 35×35 pixel (Fig.16).

Experiment 4. The instruction "Please trace on the character shape by one stroke within 5 seconds" is displayed and the challenge image Fig.10 is displayed. The small zone on the grid is 35×35 pixel (Fig.17).

Experiment 5. The instruction "Please trace on the character shape by one stroke within 5 seconds" is displayed and the challenge image Fig.10 is displayed. However, the response is accepted only when every small zone from 1 to 40 is passed in order. The small zone on the grid is 35×35 pixel (Fig.17).

Experiment 6. The instruction "Please perform the input which is not related to the displayed character" is displayed and the challenge image Fig.9 is displayed. The small zone on the grid is 35×35 pixel (Fig.16).

Table 2. Correspondence Table

	Im1	Im2	Im3	Im4	Im5	Im6
Character α	√	√	√	-	-	√
Character β	-	-	-	√	√	-
35×35 pixel	-	√	√	√	√	√
70×70 pixel	√	-	-	-	-	-
Specified time (about 5 seconds)	-	-	√	√	√	-

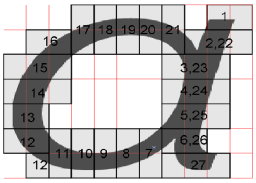


Fig. 15. exp 1

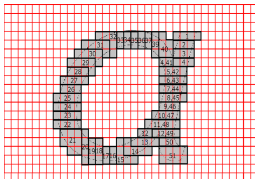


Fig. 16. exp 2-3

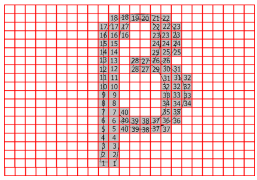


Fig. 17. exp 4-5

4.2 Result of Experiments

The results of the experiments in Section 4.1 are summarized in Table 3. In the experiment 1, 2, and 3, one subject is rejected because the responding data is in the order corresponding to the alphabet "a". Recall that the image indicates the symbol "α". When one writes "α", the order is different from the alphabet "a" although the shape is similar. The difference of the handwritten input of "α" and "a" is due to the culture and a social background in which the subject has grown up, and this is considered an embodied knowledge.

By the results of experiments 1 and 2, we can conclude that if the zone is bigger, then higher acceptance rate is achieved, on the other hand, if the zone is smaller, then acceptance rate decreased. By the results of experiments 2 and 4, we can conclude that the shape of the character does not affect the acceptance rate and the acceptance rate is stable for any (simple) character. We are convinced that other characters which are written as one-stroke sketch other than "α" can be used in the proposed CAPTCHA as well. By the result of experiment 3, we can conclude that if we allow users to write slowly then the acceptance rate will increase but the transmission data gets larger, which is not desired for network congestions. By the results of experiments 4 and 5, we can conclude that it is necessary to permit width of the order of the inputted coordinates to some degree, that is, we must be tolerant to small errors data, possibly caused by an unintentional errors. More experiments and detailed analysis will be reported in the expanded version of the paper.

Table 3. Test Result

	Test1	Test2	Test3	Test4	Test5	Test6
Number of Subjects (People)	22	22	22	22	22	22
Acceptance Number (Time)	21	20	21	22	9	0
Acceptance Rate (%)	95.5	90.9	95.5	100	40.9	0

5 Future Work and Summary

We shall discuss several issues on the proposed CAPTCHA for future research. The response data to a challenge image of the proposed CAPTCHA consists of a series of coordinates. One coordinate consists of a pointer ID, x coordinate, and y coordinate and each data is 9 bytes, where, pointer ID indicates a human action on the touchscreen. The number of input data comprises about 150-200 coordinates in our experiment using the symbol "α". One coordinate is inputted per 0.01-0.02 seconds. Therefore, 1.35-1.8 kilobytes transmission is required for each response for the proposed CAPTCHA. The response data for the existing CAPTCHA is 6-10 characters, and the transmission data is several bytes. Thus, the transmission data is bigger than the existing CAPTCHA. The proposed CAPTCHA is required more computation to check whether or not a response can be accepted than the existing CAPTCHA. We will study how to reduce amount of transmission data and server's information processing.

In our experiment we made the challenge images and the authentication programs by hand. Automatic generation of the challenge image is necessary when we use it in a real system. Because one-stroke sketch is an embodied knowledge, it is important to devise a method to put embodied knowledge in challenge images and to apply continuous transformations to a character in order not to change

the writing order. It should be note that there is no difference in the programs on Android OS but the adjustment of coordinates for platform smartphones is necessary. We will discuss these issues in the extended version of the paper.

In this paper, we propose a new CAPTCHA technique utilizing touchscreens to solve an inconvenience caused by the existing CAPTCHAs when using smartphones. We implement the proposed technique and carry out experiments to examine the usefulness and compare with the existing techniques. Using a touchscreen, one-stroke sketch is captured and represented as ordered series of coordinates. One-stroke sketch can be considered as one of embodied knowledges of human beings and so computer programs have difficulty to understand one-stroke sketch. Our technique is based on embodied knowledge of human beings and so computer programs cannot respond correctly to a challenge image. It is necessary to study more one stroke sketch as an embodied knowledge of human beings and validity and security of the proposed technique in the context of artificial intelligence and cognitive science.

References

1. von Ahn, L., Blum, M., Hopper, N., Langford, J.: CAPTCHA: Using Hard AI Problems for Security. In: Biham, E. (ed.) EUROCRYPT 2003. LNCS, vol. 2656, pp. 294–311. Springer, Heidelberg (2003)
2. von Ahn, L., Blum, M., Langford, J.: Telling Humans and Computers Apart Automatically. Communications of the ACM 47, 56–60 (2004)
3. von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science 321, 1465–1468 (2008)
4. Elson, J., Douceur, J.R., Howell, J., Saul, J.: Asirra: a CAPTCHA that Exploits Interest-Aligned Manual Image Categorization. In: ACM Conference on Computer and Communications Security, pp. 366–374 (2007)
5. Gossweiler, R., Kamvar, M., Baluja, S.: What's up CAPTCHA?: a CAPTCHA Based on Image Orientation. In: International Conference on World Wide Web, pp. 841–850 (2009)
6. Mori, G., Malik, J.: Recognizing Objects in Adversarial Clutter: Breaking a Visual CATCHA. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 134–141 (2003)
7. Turing, A.M.: Computing Machinery and Intelligence. Mind 59(236), 433–460 (1950)
8. `http://www.yamamoto-works.jp`

Code Based KPD Scheme with Full Connectivity: Deterministic Merging

Pinaki Sarkar[1] and Aritra Dhar[2]

[1] Department of Mathematics, Jadavpur University, Kolkata – 700032, India
pinakisark@gmail.com
[2] Department of Computer Science, III Technology – Delhi, New Delhi – 110020, India
aritra1204@iiitd.ac.in

Abstract. Key PreDistribution (KPD) is one of the standard key management techniques of distributing the symmetric cryptographic keys among the resource constrained nodes of a Wireless Sensor Network (WSN). To optimize the security and energy in a WSN, the nodes must possess common key(s) between themselves. However there exists KPDs like the Reed Solomon (RS) code based schemes, which lacks this property. The current work proposes a deterministic method of overcoming this hazard by merging exactly two nodes of the said KPD to form blocks. The resultant merged block network is fully connected and comparative results exhibit the improvement achieved over existing schemes. Further analysis reveal that this concept can yield larger networks with small key rings.

Keywords: Key predistribution (KPD), Reed Solomon (RS) codes, Combinatorial Designs, Deterministic Merging Blocks, Connectivity, Security.

1 Introduction

The increasing necessity of dealing with classified information from hazardous deployment area is enhancing the popularity of Wireless sensor networks (WSN). Such networks typically consists of Key Distribution Server (KDS) or Base Station (BS), identical (low cost) ordinary sensors (or nodes) and at times some special nodes. The BS links the network to the user and sometimes these networks have more than one BS. This along with other flexibilities like lack of fixed infrastructure imply that such networks are *Ad Hoc* in nature.

Each entity constituting a WSN typically consists of a power unit, a processing unit, a storage unit and a wireless transceiver. Capacities of each such unit in any ordinary node is quite limited for any WSN while the KDS is quite powerful. Resource constrained nodes are supposed to gather specific information about the surrounding, process them and communicate to the neighboring nodes, i.e., nodes within their (small) *radius of communication*. The processed data is relayed up to the KDS which has relatively (large) *radius of communication* for further analysis before informing the user.

In spite of all the weaknesses in the basic building blocks of WSNs, these networks have several military applications like monitoring enemy movements, etc. Besides they are utilized for other scientific purposes like smoke detection, wild fire detection, seismic activity monitoring etc. In all its applications, WSNs once deployed are expected to work unattended for long duration of time while its constituent nodes deals with lot of sensitive data.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 141–151, 2013.

1.1 Related Works

Most recent applications of WSNs require secure message exchange among the nodes. One ideally likes to apply lightweight symmetric key cryptographic techniques in order to avoid heavy or costly computations within the resources constraints nodes. Such cryptographic techniques demands the communicating parties to possess the same key prior to message exchange. Standard online key exchange techniques involving public parameters or using trusted authorities are generally avoided. Instead, *Key PreDistribution (KPD)* techniques are preferred.

Eschenauer and Gligor [5] suggested the pioneering idea of KPD scheme where:

- *Preloading of Keys* into the sensors prior to deployment.
- *Key establishment*: this phase consists of
  - *Shared key discovery*: establishing shared key(s) among the nodes;
  - *Path key establishment*: establishing path via other node(s) between a given pair of nodes that do not share any common key.

Random preloading of keys means that the *key rings* or *key chains* are formed randomly. In [5], *key establishment* is done using *challenge and response* technique. Schemes following similar random preloading and probabilistically establishing strategy are called *random KPD schemes*. More examples of such schemes are [4,7]. Çamptepe and Yener [2] presents an excellent survey of such schemes.

On the other hand, there exists KPD schemes based on deterministic approach, involving *Mathematical* tools. Çamptepe and Yener [1] were first to propose a deterministic KPD scheme where keys are preloaded and established using *Combinatorial Designs*. Following their initial work, numerous deterministic KPD schemes based on combinatorial designs like [6,8,10] have been proposed. There exists *hybrid* KPDs like [3,9] that use both random and deterministic techniques.

There exists some interesting designs like one in [8] using Reed Solomon (RS) code which can be viewed as a combinatorial design. One may refer to [6] for discussions about various combinatorial designs necessary for this paper. For the sake of completeness, an outline on combinatorial designs is presented in Section 3. The said section establishes that the RS code based KPD [8] can be treated as a Group–Divisible Design (GDD) or Transversal Design (TD).

1.2 Contributions in This Paper

The original scheme of [6] lacks full communication among nodes as a pair of nodes may not share a common key. This involves an intermediate node which increase the communication overhead. As a remedial strategy, Chakrabarti et al. [3] first suggested the idea of *random merging of nodes* to form blocks. Their strategy was to randomly merge 'z' nodes of [6] to form blocks having bigger key rings. The resultant network thus possessed '$\lfloor \mathscr{N}/z \rfloor$' blocks where $\mathscr{N}$ is the number of nodes in the original KPD [6]. However full communication was still not guaranteed and many aspects of their design, like the basic concept of merging, choice of nodes while merging, the heuristic in [3, Section 4] etc.have not been explained. A similar random merging concept was proposed by Sarkar and Dhar [9] for the RS code based KPD [8]. Full connectivity is

guaranteed for $z \geq 4$ (see [3, Theorem 1]) The solution to be presented here is entirely different and performs more efficiently.

Motivated by the merging concept, the present authors thought of proposing a deterministic merging technique. Here exactly two (2) nodes of the KPD [8] are merged. Theorem 2 of Section 4 establishes that merging two nodes of the KPD [8] in a certain fashion results in full communication among the newly formed (merged) blocks. The resiliency and scalability are also comparable.

2 Basics of Combinatorial Design

This section briefly describes some basic notion of *combinatorial design* necessary for understanding Ruj and Roy [8] scheme.

Group-Divisible Design (GDD) of type g^u, block size k: is a triplet $(\mathscr{X}, \mathscr{H}, \mathscr{A})$:

1. $\mathscr{X}$ is a finite set with $|\mathscr{X}| = gu$.
2. $\mathscr{H}$ is a partition of $\mathscr{X}$ into u parts, that is, $\mathscr{H} = \{H_1, H_2, H_3, \ldots, H_u\}$ with $\mathscr{X} = H_1 \cup H_2 \cup H_3 \cup \ldots \cup H_u$, $|\mathscr{H}_i| = g \; \forall \; 1 \leq i \leq u$ and $\mathscr{H}_i \cap \mathscr{H}_j = \phi \; \forall \; 1 \leq i \neq j \leq u$.
3. $\mathscr{A}$ is the collection of blocks of $\mathscr{X}$ having the following properties: $|H \cap A| \leq 1 \; \forall \; H \in \mathscr{H}$, $\forall \; A \in \mathscr{A}$, given any pair of varieties $x \in H_i, y \in H_j$ with $i \neq j \; \exists$ unique $A \in \mathscr{A}$ such that $x, y \in A$.

Transversal Designs (TD(k,n)): are special type of GDDs with $g = n, u = k$, while the parameter k remains the same. These can be shown to form $(nk, n^2, n, k)-$ configuration. One is referred to [6, Section III] for definition of configuration and other related concepts.

Common Intersection Design (CID): Let $(\mathscr{X}, \mathscr{A})$ is a (v, b, r, k)–configuration. Recall from [6], $(\mathscr{X}, \mathscr{A})$ is called a $\mu-$common intersection design (μ–CID) if: $|\{A_\alpha \in \mathscr{A} : A_i \cap A_\alpha \neq \phi \text{ and } A_j \cap A_\alpha \neq \phi\}| \geq \mu$ whenever $A_i \cap A_j = \phi$. While for the sake of consistency, in case $A_i \cap A_j \neq \phi, \forall \; i, j$ one defines $\mu = \infty$.

Maximal CID: For any given set of parametric values of (v, b, r, k), such that a configuration can be obtained with them, one would like to construct a configuration with maximum possible μ. This *maximal value of* μ is denoted μ^*. *Theorem 14.* of [6, Section IV] establishes $TD(k, n)$ designs are $k(k-1)^* - CID$.

3 KPD Using Reed Solomon (RS) Codes

This section is devoted to the description of KPD scheme proposed by Ruj and Roy in [8]. The scheme uses Reed Solomon (RS) codes to predistribute and establish the communication keys among the sensor nodes. The construction of RS codes has been given in [8]. Salient features are being sketched below:

To construct (n, q^l, d, q) RS code having alphabet in the finite field $\mathbb{F}_q$ (q: prime or prime power > 2), consider the following set of polynomials over $\mathbb{F}_q$:

$$\mathscr{P} = \{g(y) : g(y) \in \mathbb{F}_q[y] \; deg(g(y) \leq l-1\}.$$

Thus the number of elements in $\mathscr{P}$ denoted by $|\mathscr{P}| = q^l$. Let $\mathbb{F}_q^* = \{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{q-1}\}$ be the set of non-zero elements of $\mathbb{F}_q$. For each polynomial $p_m(y) \in \mathscr{P}$, Define

$cp_m = (p_m(\alpha_1), p_m(\alpha_2), \ldots, p_m(\alpha_{q-1}))$ to be the m^{th} codeword of length $n = q-1$. Let $C = \{cp_m : p_m(y) \in \mathscr{P}\}$ be the collection of all such code words formed out of the polynomials over $\mathbb{F}_q$. This results in a RS code. Since the number of code-words is q^l, the system can support up to q^l nodes.

Now the polynomial p_m and the corresponding codeword cp_m are given to the m^{th} node. For the codeword $cp_m = (a_1, a_2, \ldots, a_n)$, one assigns the keys having key-identifiers $(a_1, \alpha_1), (a_2, \alpha_2), \ldots, (a_n, \alpha_n)$ where $a_j = p_m(\alpha_j), j = 1, 2, \ldots, n$ to the m^{th} node. The node id of the m^{th} node is obtained by evaluating the polynomial p_m at $x = q$ and taking only the numerical value. That is the m^{th} node has the node id $p_m(q)$ (without going modulo 'p').

A WSN with 16 nodes based on RS code parameters $q = 4, n = 3$ and $l = 2$ is presented in Table 1. Here '2' means the polynomial 'x' and '3' means the polynomial '$x+1$' modulo the irreducible polynomial $x^2 + x + 1$ over $\mathbb{F}_2[x]$ which are commonly referred to as $\overline{x}$ and $\overline{x+1}$. Thus $0, 1, 2, 3$ forms the finite field $\mathbb{F}_4$. The nodes' polynomials $i + jy \in \mathbb{F}_4[y]$ for $0 \le i, j \le 3$ are given in 2nd row of Table 1. By evaluating these polynomials at non-zero points, the keys $(p_m(b), b), 0 \ne b \in \mathbb{F}_q, 0 \le i, j \le 3$ have been derived and tabulated in the corresponding columns.

Table 1 constructed by similar computations is being presented in a slightly different manner from Ruj and Roy [8]. This *GDD* form of presentation helps one realize the similarity of the RS code based KPD of [8] with the $TD(q-1, q)$ with parameters $q-1, q$ of [6]. Though in Theorem 6 of [6, Section III], constructions $TD(k, p), 2 \le k \le q, p$ a prime is given, it can be extended to $TD(k, q), q = p^r$. Since the construction of the KPDs $TD(k, p)$ of [6] utilized the field properties of $\mathbb{F}_p$, one can extend it to $\mathbb{F}_q = \mathbb{F}_{p^r}$. Extending the base field from $\mathbb{F}_p$ to $\mathbb{F}_q = \mathbb{F}_{p^r}$ and following similar constructions as given in [6, Section III] yields $TD(k, q), q = p^r, r \in \mathbb{N}$. Now taking $k = q-1$ results in $TD(q-1, q)$. However it is important to state that in $TD(q-1, q)$ design is different from RS code. In $TD(q-1, q)$ design, the evaluation is done for $y = 0, 1, \ldots, q-2$ while in RS code based design, it is done at non–zero points, $y = 1, 2, 3, \ldots, q-1$.

N_0 to N_{15} denotes the nodes with ids ranging from 0 to 15 whose polynomials are represented in the column immediately below it. Key ids contained in a node are presented in the columns below each node. V/C denotes the distinct *Variety Classes* H_1, H_2, H_3, where $H_d = \{(i, d) : 0 \le i \le 3\}$ for $d = 1, 2, 3$. One notes that the scheme under consideration is a $(q-1)(q-2)-$CID as the number of keys per node $= k = q-1$ (see Section 2). Thus for nodes not sharing any key, there are enough nodes which can play the role of the intermediate node in multi-hop (2-hop) process. This encourages the search for a deterministic merging design with exactly two nodes per block yielding full communication among the blocks.

3.1 Weakness of the Above RS Code Based KPD

Apart from other possible weaknesses, the RS code based KPD presented in [8] lacks full communication among nodes (by the discussions above). So multi-hop communications occur among the nodes. Here multi-hop means some third party node other than the sender and receiver decrypts and encrypts the ciphertext. Other than increasing the cost of communication, this enhances the chances of adversarial attacks on such

communication. Thus the energy efficiency as well as the security of message exchange of the entire network might be grossly affected.

4 Remedy: Deterministic Merging of Nodes

Lack of direct communication for any arbitrarily chosen pair of nodes of the KPD [8] can be tackled by merging certain number of nodes. For this, observe that Table 1 indicates the network having 16 nodes can be partitioned into 4 classes each containing 4 nodes on the basis of their key sharing. These classes are separated by double column partitioning lines after each set of 4 nodes: N_0, N_1, N_2, N_3; N_4, N_5, N_6, N_7; N_8, N_9, N_{10}, N_{11}; and $N_{12}, N_{13}, N_{14}, N_{15}$. Every class has the property that the coefficient of y in their respective polynomials is same. Equating each other's polynomials $i + jy$ with $0 \leq i \leq 3$ for some fixed $j = 0, 1, 2$ or 3 results in no common solution $\implies$ no common key. For eg. with $j = 1$ with $i = 0, 1, 2, 3$, the corresponding 4 polynomials: $0 + y, 1 + y, 2 + y, 3 + y$ do not have any common solution. Hence no shared keys for corresponding nodes.

Table 1. Polynomials, Node and Key identifiers for $q^2 = 16$ nodes. Table adapted from section 3.1 of Ruj and Roy [8]. Alternative presentation: Group–Divisible Design (GDD) form.

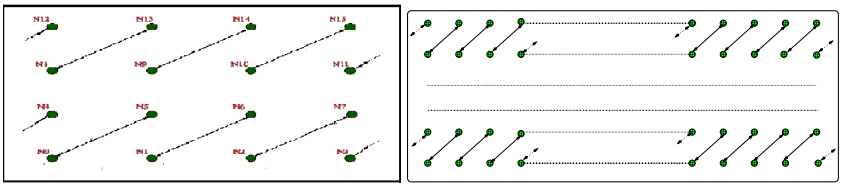
Nodes	N_0	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	N_9	N_{10}	N_{11}	N_{12}	N_{13}	N_{14}	N_{15}
V/C	$0y+0$	1	2	3	y	$y+1$	$y+2$	$y+3$	$2y$	$2y+1$	$2y+2$	$2y+3$	$3y$	$3y+1$	$3y+2$	$3y+3$
H_1	(0,1)	(1,1)	(2,1)	(3,1)	(1,1)	(0,1)	(3,1)	(2,1)	(2,1)	(3,1)	(0,1)	(1,1)	(3,1)	(2,1)	(1,1)	(0,1)
H_2	(0,2)	(1,2)	(2,2)	(3,2)	(2,2)	(3,2)	(0,2)	(1,2)	(3,2)	(2,2)	(1,2)	(0,2)	(1,2)	(0,2)	(3,2)	(2,2)
H_3	(0,3)	(1,3)	(2,3)	(3,3)	(3,3)	(2,3)	(1,3)	(0,3)	(1,3)	(0,3)	(3,3)	(2,3)	(2,3)	(3,3)	(0,3)	(1,3)

The other case for any pair of nodes not sharing any common key is whenever their constant term is same since only non–zero values of y are allowed. This gives rise to alternative type of partition: N_0, N_4, N_8, N_{12}; N_1, N_5, N_9, N_{13}; N_2, N_6, N_{10}, N_{14}; and N_3, N_7, N_{11}, N_{15}. This motivates one to visualize the key sharing of the 16 nodes, N_0 to N_{15}, like a 'square-grid' as presented in Figure 1(a). Any pair of nodes, other than the ones lying in the same row or column shares exactly 1 key as the equations: $(j - j')y = (i' - i)$ has unique solution over non–zero points of $\mathbb{F}_4$ (since $q = 4$) that is with $0 \leq i \neq i', j \neq j' \leq 3$. Merging of nodes in pairs for the case $q = 4$ can be now achieved as indicated by the slanted line in Figure 1(a). Basically the strategy is to merge the nodes: $N_{(i,j)}$ and $N_{(i\oplus 1, j\oplus 1)}$ where $\oplus$: addition in $\mathbb{F}_4$ (addition modulo 2), for $j = 0, 2$.

A natural deterministic merging strategy of 2 nodes can now be visualized for $q = 2^r \implies \mathscr{N} = q^2 = 2^{2r}$. Figures 1(b) demonstrate the strategy. Nodes occurring at the ends of a slanted line are merged. Idea is to break up the network into pairs of rows, i.e. $\{1,2\}; \{3,4\}, \ldots, \{2^{r-1}, 2^r\}$ and apply similar process.

Before explaining the general odd case, it is useful to visualize the case when $q = 5$, i.e. a network with $q^2 = 5^2 = 25$ nodes as presented in Figure 2(a). Rest of the discussion is similar to that of the case $q = 4 = 2^2$ except for the merging of last three rows. As usual the arrows indicate the merging strategy. The strategy indicated in Figure 2(a) is same for the first and second rows while differs in the last three rows when compared to

Figure 1(a). This is because a similar strategy like previous cases would imply one row is left out. Of course for $q = p = 5$ all arithmetic operations are 'modulo 5' arithmetic operations as done in $\mathbb{F}_5$.



(a) MB Design 1: Particular case $q = 4 = 2^2 \implies \mathscr{N} = 16 = 4^2$.

(b) MB Design 1: General even case for $q = 2^r, r \in \mathbb{N} \implies \mathscr{N} = 2^{2r}$ nodes

Fig. 1. Deterministic Merging Blocks Strategy for all even prime power cases $q = 2^r, r \in \mathbb{N}$

For general odd prime power case $q = p^r : p$ is a odd prime, any pair of nodes among $\{N_{(i,j)} : 0 \leq j \leq q-1\}$ for every $0 \leq i \leq q-1$ (i.e., nodes with ids $i + jq$ and occurring in i^{th} row; $q-$fixed) do not share any common key. Same is the case with the nodes $\{N_{m,j} : 0 \leq m \leq q-1\}$ *for every* $0 \leq j \leq q-1$ with ids $m + jq$ (occurring in j^{th} column of Figure 2(b)). For any other pair of nodes, equating corresponding linear polynomials they find exactly one (1) common shared key between them (since $l = 2$). The general case of q^2 nodes ($l = 2$) can visualized as a $q \times q$ 'square-grid' as in Figure 2(b) with nodes in same row or column having no key in common while any other pair of nodes share exactly one common key. This merging strategy is indicated in Figure 2(b) by slanted arrows as before. Nodes occurring at the ends of a slanted line are merged. The idea is to look at two rows at a time and form blocks containing one node of each except for last 3 rows. For fixed $0 \leq i \leq q-2$, merge the nodes $N_{(i,j)}$ and $N_{(i\oplus 1, j\oplus 1)}(\oplus :$ addition modulo $q)$, for $0 \leq j \leq q-3$ (with increment of 2) $\forall\, q > 4$. The last node of every odd row is merged with the first node of the row above it. Since q is odd, taking combination of two row would have left out one row, so top three row are combined separately.

Note that, in case merging of nodes is done randomly, one may end up with merged pairs like $N_{(0,0)} \cup N_{(0,1)}$ and $N_{(0,2)} \cup N_{(0,3)}$, (for $q \geq 4$) which do not share any common key, thus not be able to communicate even after merging.

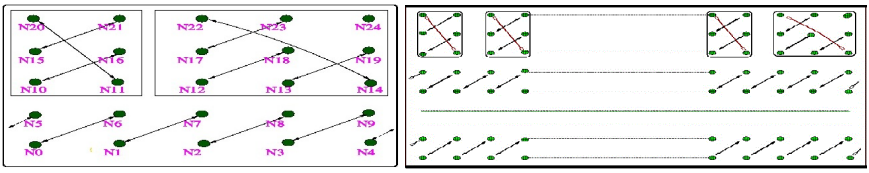
4.1 Assured Full Communication: Theoretical Results

Equating the polynomials of the 4 nodes constituting any 2 merged blocks yields:

Theorem 1. *The proposed Deterministic Merging Block Strategy where 2 nodes of the RS code based KPD [8] are clubbed to form the merged blocks results in full communication among the merged blocks.*

Proof. Consider any two arbitrary blocks A and B. It is evident from the construction that at least node from block A will never lie in the horizontal line as well as the vertical

line of either of the two nodes the other block B (refer to Figures 1(a), 1(b) 2(a) and 2(b) for $q = 4, 2^r, 5$ and for general case). This implies that these two nodes will have a common key as discussed in Section 4. Hence the blocks A and B can communicate through this key. As the two blocks were arbitrarily chosen, one is assured of full communication in the new network consisting of blocks constructed by merging two nodes in the manner explained above (see Figures 1(a), 1(b) 2(a) and 2(b) for $q = 4, 2^r, 5$ and for general case respectively).



(a) MB Design 1: Special case for $q = p = 5 \implies \mathcal{N} = 25 = 5^2$ nodes.

(b) MB Design 1: General odd prime / prime power $q = p^r, r \in \mathbb{N} \implies q^2$ nodes.

Fig. 2. Deterministic Merging Blocks Strategy for all odd prime power cases $q = 2^r, r \in \mathbb{N}$

Theorem 2. *The resulting Merged Block Design has a minimum of* one *to a maximum of* four *common keys between any two given pair of (merged) blocks.*

Proof. Any two nodes can share at most one key in original RS code based KPD in [8]. So there are at most 4 keys common between two blocks. This situation occurs only if both nodes of the 1^{st} block shares two (2) distinct keys with each node of the 2^{nd} block.

Remark 1. Some important features of the merging block design are as follows:

- Merging does not mean that the two nodes combine physically to become one. Just that they are to be treated as one unit.
- The resultant merged block design has full communication among the blocks through at least one common key between any two given pair of merged block ensuring full communication in resultant network.
- Full communication can not be assured when nodes are merged randomly to form larger blocks. Probably this is the main reason why authors of [3] could not justify several issues in their random merging model.
- The current authors feel that it is mandatory to have inter nodal communication. Any communication received by either of the two constituent nodes of a block can be passed down to the other node and hence make the other node connected. As such, while proposing the merged block design, this consideration was given importance.
- Therefore the total number links in the merged scheme is same as that of the original RS code based KPD of Ruj and Roy in [8]. This fact will be recalled later while discussing resiliency of the system.

5 Network Parameters

Some important aspects of the combined scheme like *communication probability, computational overhead, resiliency and scalability* will be present in this section.

5.1 Communication Probability; Overhead; Network Scalability

Communication Probability or *Connectivity* is defined to be the ratio of number of links existing in the network with respect to the total number of possible links. A *link* is said to exists between two nodes they share at least one common key. Figure 3 presents a comparison between the present scheme v/s existing schemes in terms of $(E(s))$ and connectivity. The graph in Figure 3(a) is plotted with number of nodes in the network in x-axis v/s $E(s)$ in y-axis. It compares the resiliency of Merged Block network with other existing schemes. The graph plotting in Figure 3(b) based on the varying number of keys per node, k in x-axis) v/s connectivity in y-axis. The original KPD [8] is assumed to have $q = 49, \mathcal{N} = 2401$ many nodes. So that the merged network has $\mathcal{N}_{mb} = 1200$ nodes. Clearly the present Merging Block Design provides much better communication than the original RS code based KPD [8] and the random models of [3,9] (both almost same) even when no. of keys per nodes, (k) decreases. For the RS code based KPD [8], taking the key ring of node (i, j) as $\{(p_{i+jq}(\alpha_c), \alpha_c) : h \leq \alpha_c \leq q-1\}, 1 \leq h \leq q-2$ yields decreasing key rings for increasing h. The present Merging Block Design over [8] possesses *full connectivity* $\forall\, k \geq \lfloor \frac{q+7}{4} \rfloor$. This follows from the observations that any 2 block share a min. of $4k-6$ keys, there are q keys in the network and the *pigeon-hole-principle*. Present design connectivity $= 1\ \forall\ k \geq 14$ as $q = 49$.

Communication overhead measures the computational complexity of both the key establishment and the message exchange protocols. During *key establishment* polynomials of $(l-1)^{th}$ degree are equated which involves computing inverses over $\mathbb{F}_q$. Since quadratic, cubic and quartic equations can be solved in constant time, this design is valid for for $l = 2, 3, 4$ and 5. However complexity of quadratic, cubic and quartic is quite high, specially for the nodes. Even the resiliency falls drastically with increasing value of l. So practically $l = 2$ is considered. For message exchange, complexity is same as that of original RS code based KPD [8]. Of course the complexity depends on the cryptosystem being used. Any *cryptosytem* meant for *embedded systems like AES-128*, is applicable.

5.2 Resiliency

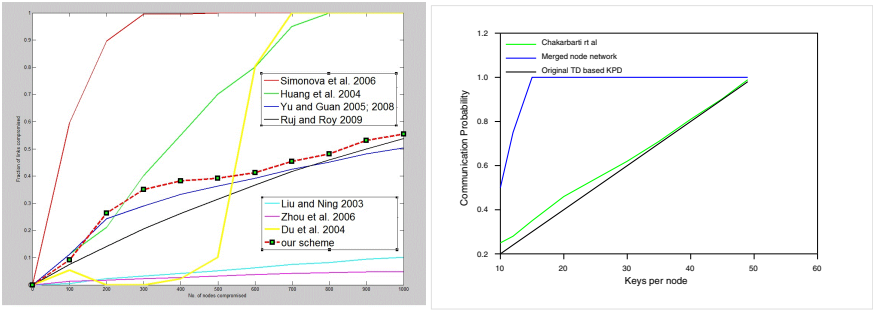
Before proceeding with further analysis in the remaining part of the paper, some terminologies need to be introduced. The term 'uncompromised node(s)' associates to node(s) that are not compromised/captured. The *link* between any two uncompromised nodes is said to be *disconnected* if all their shared key(s) gets exposed due to capture of s nodes. Standard resiliency measure, $E(s)$ (recalled below) is considered while analyzing/comparing the scheme with existing KPDs.

$E(s)$ measures the ratio of number of links disconnected due to capture of s nodes (here blocks) with respect to the total number of links in the original setup. Mathematically: $E(s) = \frac{\text{number of links broken due to capture of s nodes (here blocks)}}{\text{total number of link in the original setup}}$

One can refer to [8, Section 6.4] for an estimated upper bound of $E(s)$. Construction of the Merging Blocks clearly indicated that a merged network of size $\mathcal{N}$ corresponds to $\approx 2\mathcal{N}$ sized original KPD, while capture of s merged blocks is almost equivalent $2s$ nodes of original. Thus the ratio of links broken ($E(s)$) remains almost the same ($\approx$ old $E(s)$) as observed experimentally. Providing accurate theoretical explanations is rather difficult due to the randomness of node capture both the original and its merged KPDs.

6 Simulation and Comparative Results

Run results after 100 runs for each data set are tabulated in Table 2. $\mathcal{N}_{RS}(=q^2)$ denotes the number of nodes of the original KPD scheme in [8]. While $\mathcal{N}_{MB}$ denotes the



(a) Comparative E(s) values. (b) Comparative connectivity values.

Fig. 3. Graphs showing the comparison between the MB design over RS code based KPD v/s existing schemes with regards to connectivity and resiliency (E(s)) values

Table 2. Simulated $E(s)$ results for MB over RS code KPD: Comparison with RS code KPD

$k = q-1$	$\mathcal{N}_{RS}$	$\mathcal{N}_{MB}$	s_{MB}	s_{RS}	$E_{MB}(s)$	$E_{RS}(s)$
28	841	420	5	10	0.297007	0.297155
28	841	420	10	20	0.507596	0.509076
30	961	430	5	10	0.279395	0.280538
30	961	430	10	20	0.483049	0.484641
40	1681	840	5	10	0.219206	0.219512
40	1681	840	10	20	0.390549	0.391530
48	2401	1200	5	10	0.186518	0.186756
48	2401	1200	10	20	0.338435	0.338898
70	5041	2570	10	20	0.247447	0.247348
70	5041	2570	15	30	0.346970	0.347509
100	10201	5100	5	10	0.094725	0.094747
100	10201	5100	10	20	0.180525	0.180588
100	10201	5100	20	40	0.328717	0.328873
102	10609	5304	10	20	0.177422	0.177433
102	10609	5304	30	60	0.443885	0.444061

number of blocks in merged network. Clearly $\mathscr{N}_{MB} = \lfloor \frac{\mathscr{N}_{RS}}{2} \rfloor$. Here, p is a prime number and $q = p^r$ is a prime power. Any given pair of nodes has at most 1 key in common as $l = 2$. Let s_{MB} and s_{RS} be the number of blocks and nodes captured in the original and its merged KPD respectively. Then $E_{RS}(s)$ and $E_{MB}(s)$ denotes the resiliency coefficients in the original and merged blocks respectively. The mentioned tables compares the simulated values of ratio of links disconnected $E(s)$ in the Merging Block model with its original KPD [8].

7 Conclusions and Future Works

Deterministic merging of nodes of RS code based KPD [8] has been proposed in this paper. As a result full communication between nodes is achieved. Though the merging is being done for the particular chosen KPD scheme in [8], the approach can be modified and generalized to other schemes. This enhances the applicability of the concept. To understand why deterministic is better than its random counterpart, the mentioned design is viewed combinatorially. Algebraic as well as design theoretic analysis of the mentioned KPD paves the logical approach behind the deterministic merging strategy. Remark 1 of Section 4.1 highlight some of the important features of the deterministic merging strategy.

One can readily visualize some immediate future research directions. Application of similar deterministic merging concept to network based on other KPDs which lacks full communication among its nodes like may result in interesting works. The reason of preferring such merging strategy over its random counterpart has been sketched. Deterministic approach can be generalized to other schemes like [6]. Generic survey of deterministic v/s random schemes (like current scheme v/s [3,9]) yielding fully communicating networks can be future research topics.

The assurance of connectivity with lower keys (refer in Figure 3(b)) paves a direction of achieving fully communicating deterministic schemes having high resiliency. A priori one must look to design scheme having good node support, small key rings, high resilience and scalability. Mathematical solutions to such fascinating problems will be interesting. The deterministic property of the merging technique may enable it to be combined with other deterministic techniques like the one proposed in [10]. This will ensure that the merged design is free of 'selective node attack' which it still suffers from as the original KPD did.

References

1. Çamtepe, S.A., Yener, B.: Combinatorial Design of Key Distribution Mechanisms for Wireless Sensor Networks. In: Samarati, P., Ryan, P.Y.A., Gollmann, D., Molva, R. (eds.) ESORICS 2004. LNCS, vol. 3193, pp. 293–308. Springer, Heidelberg (2004)
2. Çamtepe, S.A., Yener, B.: Key distribution mechanisms for wireless sensor networks: A survey 2005. Technical Report, TR-05-07 Rensselaer Polytechnic Institute, Computer Science Department (March 2005)
3. Chakrabarti, D., Maitra, S., Roy, B.: A key pre-distribution scheme for wireless sensor networks: merging blocks in combinatorial design. International Journal of Information Security 5(2), 105–114 (2006)

4. Chan, H., Perrig, A., Song, D.X.: Random key predistribution schemes for sensor networks. In: IEEE Symposium on Security and Privacy. IEEE Computer Society, Los Alamitos (2003)
5. Eschenauer, L., Gligor, V.D.: A key-management scheme for distributed sensor networks. In: ACM Conference on Computer and Communications Security, pp. 41–47 (2002)
6. Lee, J.Y., Stinson, D.R.: A combinatorial approach to key predistribution for distributed sensor networks. In: IEEE Wireless Communications and Networking Conference, WCNC 2005, New Orleans, LA, USA (2005)
7. Liu, D., Ning, P.: Establishing pairwise keys in distributed sensor networks. In: ACM Conference on Computer and Communications Security, pp. 52–61. ACM, New York (2003)
8. Ruj, S., Roy, B.: Key Predistribution Schemes Using Codes in Wireless Sensor Networks. In: Yung, M., Liu, P., Lin, D. (eds.) Inscrypt 2008. LNCS, vol. 5487, pp. 275–288. Springer, Heidelberg (2009)
9. Sarkar, P., Dhar, A.: Assured Full Communication by Merging Blocks Randomly in Wireless Sensor Networks based on Key Predistribution Scheme using RS code. International Journal of Network Security and Its Applications, IJNSA 3(5), 203–215 (2011)
10. Sarkar, P., Saha, A., Chowdhury, M.U.: Secure Connectivity Model in Wireless Sensor Networks Using First Order Reed-Muller Codes. In: MASS 2010, pp. 507–512 (2010)

Indonesian Digital Natives

ICT Usage Pattern Study across Different Age Groups

Neila Ramdhani and Wisnu Wiradhany

Universitas Gadjah Mada

Abstract. Since its first appearance on early 2000's at the U.S, the idea that a new generation of students called digital natives or net generation has entered the world has been widely discussed by parents and educators alike. It is said that this generation think, socialize, and act differently; and they will alter roles and regulation of work and educational institutes. Now, at the second decade of the 21st century, Indonesia has to ready herself to meet this new generation. In this paper, we compared information and technology (ICT) access, activities within ICT, investment on ICT, and attitude towards ICT between five hundred Indonesian in three different groups: those who born before the 1980s; those who born between 1980s to 1989's, and those who born after the 1990s by ANOVA. We found that there were no difference on information and technology (ICT) access, activities, investment on ICT, and attitude towards ICT between the groups.

Keywords: ICT, Internet, Digital Natives.

1 Introduction

Since its emergence in the late 1970s, the internet has growth rapidly and change people's way of life [13]. At the 1980s, electronic mail has complimented face to face communication and correspondence communication. World Wide Web (web browser tool) made its debut in 1991, followed by the emergence with browser some time afterwards. At the end of the 1990s, search engines, portals, and e-commerce started to emerge. These innovations in the field of information technology are increasing at the 2000'swhen social networking sites and blogs were launched. In late 2008, music and video are integrated into the network via iTunes and YouTube [13].

Internet have changed the way people interact by eliminating physical boundaries and changing the way people manipulate the external world by opening a range of possibilities in using collaborative media. In this environment a generation who no longer intimidated by the presence of new digital technologies was born. They are able to use the technology better than their parent. Tapscott [17] extremely states that this digital generation is breathing with the technology.

In the United States, this generation is made up of citizens who were born after 1980s. They spent their teenage years in the late 1990s, where the Internet and personal computers can be found in every private home. This digital generation is

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 152–161, 2013.

known also known as net generation [17] or digital natives [13, 14]. As a generation who are familiar with technology, digital natives have different characteristics than the previous generation. Through nGenera survey from 6,000 digital generations, Tapscott [17] summarizes these differences into eight main characteristics that reduced further to four characteristics by Leung [9] as follows:

1. Global orientation and emotionally uninhibited: this generation finds it easier to share their thoughts online to everyone they can find, protected by the anonymous nature of the internet.
2. Right to information and learning: this generation beliefs in equal opportunity and freedom of expressions of opinions. Keeping up with newest information helps them to ensure their employability
3. Technological savvy: this generation has a keen interest not only in new technologies, but also how they work.
4. Preoccupied with maturity and adulthood: this generation expect to be treated as an adult, no matter how old they are.

Digital natives also use technology in different ways from their previous generations. As the generation that was born and grew up with technology, they are able to use technology in a more diverse and more efficient way than their previous generation who dealt with technology in older age. A survey from Kvavik [8] shows how digital natives use technologies for wide arrays of activities, from productive activities such as writing documents and sending e-mails to recreational activities such as random browsing and chatting with friends and colleagues. Their activities also varied from using basic technology such as word processing and presentation to using advanced technologies such as 3D Graphics and web design. These characteristics and competencies are different from previous generations in which work and entertainment are more limited, and most of the work can takes a long time in the settlement process.

If Prensky [14] calls this digital generation “digital natives,” the previous generation is called with a digital immigrant. The digital natives born and raised in the technology, but digital immigrants born before the technology was rampant. The digital natives became familiar with the technology at their adolescence, while the digital immigrant met these technologies in adulthood. With these different environment, comes different characteristics between the generations.

In the context of work for example, it can already be found digital natives who are smarter than the digital immigrants on operating software. The digital natives were able to find the information more quickly, and could complete certain work more efficiently with the help of the internet [13]. In the context of education, students were found correcting information from their teachers, with the help of information found in the internet by smart phones [17].

In Indonesia, the internet were started to use widely by the end of 1996 [4]. In contrast to the characteristics of Internet used in the United States, where each house has its own connection, internet connection in Indonesia can be find communally in internet cafes, or commonly known as warnet [4, 10] which has different characteristics from private pay system

Although having different way to access the Internet, interestingly Indonesian people have similar pattern in terms of the diversity of activities done online with Western society. Such similarity depicted in a survey conducted by Wiradhany [18], which shown that the Indonesian people are familiar with common functions such as internet browsing (surfing), chat room, and e-mail. Set of studies summarized by Hill and Sen [5] also showed that in general, Indonesia's young generation is using the internet. Campus ranks third as a place to access the internet; Undergraduate and High School Students ranked first and second levels of internet users; and students and youth as the two largest Internet user groups [5].

The goal of this paper is to identify ICT usage pattern of young internet user of Indonesia. From the previous findings, it is expected that it will show similar pattern with findings from Western Countries. Secondly, by comparing young internet users to their previous generation, this paper also wants to look up how much differently young internet user use technology compared to their previous generation.

The results of this research can enrich the discourse on the impact of technology to the people of Indonesia, since a similar study has not been done much. Psychological discourse of the digital generation can be widely diverse, ranging from discussion of internet addiction [19], the use of technology in the learning process [12], the dynamics of interaction patterns in the family [16], organizations that can maximize the potential of this generation [17], and various other areas of Psychology in contact with the digital generation. The results of this study can be used as a basis for the study of digital natives in Indonesia in the future.

2 Methods

There are three variables in this study: age, ICT-related behavior, and ICT-related attitudes. Age were grouped by those who were born in the 1970s and before, those who born between the 1980 to the 1989, and those who born after 1990s. ICT-related behavior was described by (1) accessibility and gadgets owned (2) activities within the internet and gadgets and (3) the investment for technologies monthly.

The data in this study were obtained through a survey on technology ownership, internet access, use of internet, mobile phone use, computer use, use of game consoles, and attitudinal scales of the digital generation. Items in the general survey were developed based on Wiradhany [18] findings, while the scale of the digital generation attributes were adopted from a similar scale made by Leung [9].The digital generation attitudinal scale consists of four dimensions, namely (1) the right to information, (2) unhibition in the net, (3) technological savvy, and (4) adulthood and confident. The overall scale of the attribute consists of 24 items with an internal consistency of each of the items ranged from 0.25 to 0.47 and Cronbach Alpha reliability of 0.82.

Ninety-nine percent of respondents (304 people) write their gender. Of that quantity, 56.4% (172 people) of respondents are female, while the remainders (132 people) are men. They were born in the 1970s, 1980s, and 1990s. The digital generation was assumed born in 1990s. Generation born in the two decades prior is used as a comparison to see the contrast between the use of technology between digital generation and previous generations.

Table 1. Age of Respondents

Decades	Respondents
<1979	7
1980-1989	83
≥1990	199
Total	289

Ninety-four percent of the respondents (289) wrote their date of birth. The mean of respondent's age was 21.96 with 4.26 SD. The youngest respondent is 13 years old and the oldest respondent is 60 years old by 2012.

3 Result

Survey analyses were divided into two parts, namely (A) mapping of the various technologies used and (B) attitude toward technology. In the first part, the results of this survey will be presented to the (1) Variety of technologies used across generation, (2) the activities within Internet, and (3) The amount of the investments made to access the technology.

3.1 Mapping of the Various Technologies Used

1. Variety of technologies used across the generations
Figure 1 shows the difference in technology products ownership between digital generation and the previous generation. For those who were born in the 1970s and 1980s, USB flash drive (100% and 94%) are the most widely owned product, followed by mobile phones (85.7% and 92.8%), and Internet access/laptops (71.4% and 90.4%). In the group that was born in 1990s, three top technologies are a cell phone (95%), laptops (92.5%), and Internet access (90.5%).

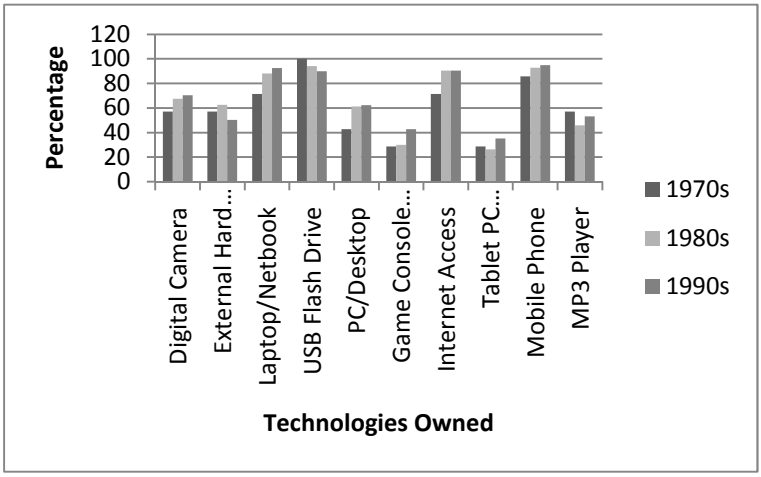


Fig. 1. Technology Owned across Generation

Data management (flash drive and laptops) and communication (mobile phone and internet) are agreed by whole groups as essentials technologies. High usage of laptops than PC, and mobile phone also shows how important mobility is. Our data from internet access also shows that internet hotspot, HSDPA modem, and mobile internet plays significant role for people to access internet.

2. Activities within Technologies

Efforts to understand the intensity and engagement of digital generation with technology were done by comparing activities within technology with the previous generation. Activities compared included: activities on mobile phone, PC/laptop, and internet. While all groups use mobile phone to send text and make phone calls, respondent born at the 1990s also used them for social networking and chat. If mobile phone is used mainly for organizing schedule, computer is mainly used to work on documents related to schoolwork or office.

All groups of respondents used computer more to access the Internet (100%, 96.4%, and 94%, respectively). Figure 2 shows that all groups go online mainly for checking their e-mails, browsing, downloading data, and social networking. It is interesting that younger people tend to do more recreational browsing while older generation browses for task-related activity.

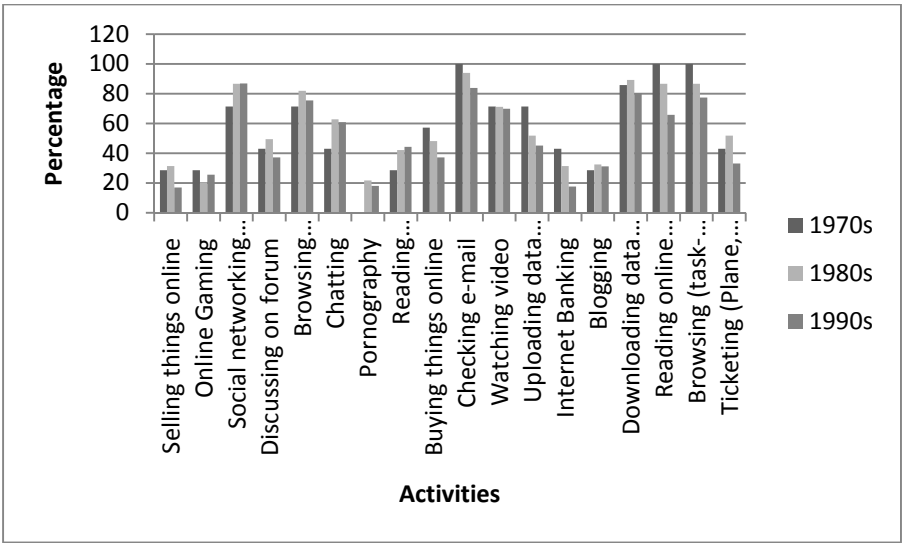


Fig. 2. Online Activities

3. Investing on Information Technology

Access to information technology in Indonesia is still a serious problem for its users. However, the price is often defeated by necessity. For the digital generation, information technology has become imperative. This survey shows that respondents born 1980s are the most likely group to spend money on the Internet than respondents born

1990s and 1970s. On average, respondents who were born in the 1970s to spend Rp. 57,857.14 per week for a mobile telephone, respondents were born in 1980s for Rp. 43,939.76 - and 1990s Rp. 40,055.13.

Another interesting investment observed to understand the information technology-related behavior is in terms of time. The data obtained showed that the range of time spent by respondents from each generation tend to be different. The respondents were born in the 1970s spent most of the time with their cell phones for non-communication activities such as listening to a song and taking pictures (31.67 minutes) and check e-mail. Those born in the 1980s spent more time on non-communication activities (25.38 minutes) and also for chatting (22.06 minutes). Diversity is becoming increasingly different on respondents born in 1990s because they mostly use his mobile phone to chat (56.42 minutes) and send short messages (33.10 minutes).

Chatting is a synchronous communication activity between two or more users of information technology. This mobile phones facilitated communication is widely used by the digital generation. Compared with communication via e-mail, chat allowed users of Internet technology to obtain near real-time response. Chat usage depicts speed as one of the characteristics of the digital generation.Speed is required by the digital generation not only in terms of obtaining the response, but also in fulfilling their needs of information. Browsing or surfing the internet is an activity done to meet these needs.

For activities within computer, the most time-consuming activities are web design, browsing, using Ms. Office, and offline gaming. No perceived differences between groups were found here. But interestingly, when asked about online activities, we found that respondents born on 1970s spent most of their time online for gaming and chat, while younger generation tend to do more browsing for task-related activities. Activities done to develop relationships that have been established through chat and online gaming seem to be two examples of activities done by1970s generation. Prensky [14] refer to this generation who born on the1970s as digital immigrants because the internet was present at the time they are adult. They are required to live and conform to the pattern of life in the era of digital technology. They treat technology as a tool to facilitate the achievement of their goals and enrich the quality of communication that had been established.

In the country with a collectivist culture, social life is managed by growing need among individuals and another. Greeting each other is not just lip service but rather form of fulfillment of needs. Media that facilitates the echo-reply brief communication like chat become a very popular choice in Indonesia. Ramdhani [15] found that in the Paperless Office communication media, social topic is the most frequent topic discussed among university lecturers.

Different trend is shown by the generation born at the 1990s. They spent more time browsing for task-related activities. Internet technology already exist at the time of this generation hit adolescence, so doing job/school related activities are more important for them.

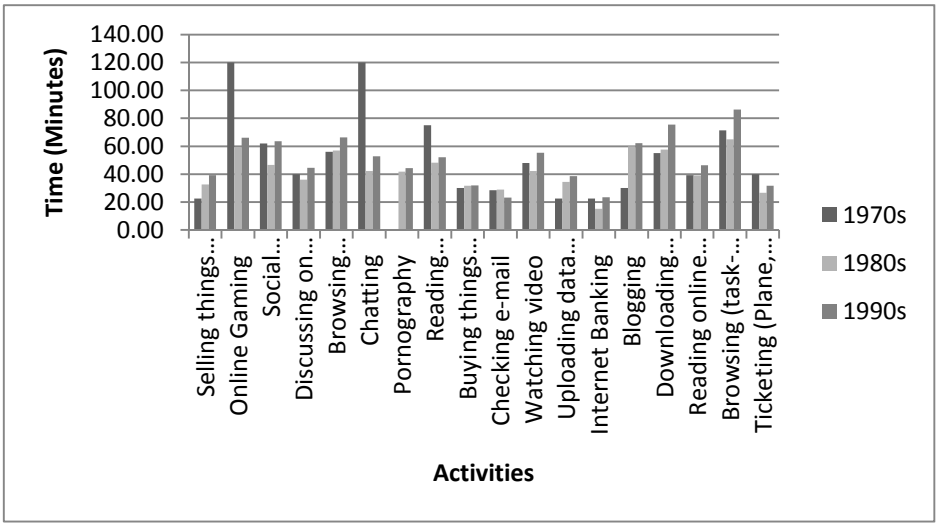


Fig. 3. Time Spent Online

3.2 Attitude towards Technology

Utilization of various technologies for life is determined by the perception of how important each feature provided by these technologies. All generations put internet and mobile phones as the technology they need. The difference between them lies in the management of data. If the generation born 1970s requires a USB Flash Drive, those who were born in the 1990sand 1980s find laptops more important. There are at least three possibilities that could be used to understand this. First, laptop with various specification and prices can be easily found in market nowadays. Second, USB Flash drives can only be used for storing data, while laptop can be used for various purposes. Third, digital generation can utilize various features of the Internet to access their files (e.g. cloud storage), so that files can be accessed from anywhere without having to be burdened with the risks of the use of USB Flash Drive.

In addition to the popular range of technology, the survey also shows that the Game Console and tablet PCs including the iPad and the Galaxy tab is a technology that is considered less important than other technologies such as laptops and computers. This may be based on the assumption that console and PC games do not directly support productivity. However, Figure 4 also shows that music player is pretty much used by all three generation.

Four attributes of the digital generation [9] is elaborated by comparing the pattern of the four attributes among respondents who were born in the 1970s, 1980s, and 1990s. The data show that the distribution of scores of all respondent groups did not differ, with the exception of right to information. Younger people tend to demand higher right to information as the number of information stored online increases [13]. However, contrary to the assumption of net generation, those born in the year 1980s has a higher net-generation attributes average score (70.18) than those born at the

1990s (68.79). The figure also shown that adulthood and confidence score has a positive correlation with age. This is expected because people tend to be more confident as they become older.

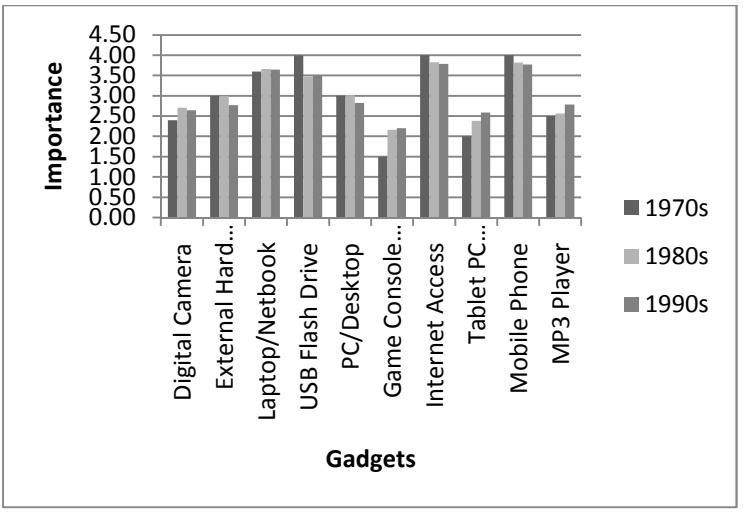


Fig. 4. Importance of Gadgets (1=Very Unimportant; 4=Very Important)

Finally, we weight and compare all of our data regarding the five attributes related to technology (access, activities, cost, time, and attitude) with each other. We then compute their differences using ANOVA.

Table 2. Net Generation Model Comparison (N = 289)

		Access	Activity	Cost	Time Spent	Attitude
Total	Mean	3.260	3.372	3.264	3.272	2.876
	Std. Deviation	1.186	0.842	2.836	2.513	0.428
	F	0.961	0.874	1.476	0.357	0.086
	Sig.*	0.384	0.418	0.230	0.700	0.917

* Significant for $p<0.05$.

From the comparison shown on table 5 it can be seen that there are no significant difference between access, activities, cost, time, and technology-related attitudes among those who were born in the 1970s, 1980s, and 1990s. This result is contrary to the first survey was done by the Americans.

4 Conclusion

While showing certain characteristics of digital generation, this survey find that those assumed as digital generation in Indonesia show no different pattern in technological access, activity, cost, time spent, and attitude from their previous two generations. These results are in line with Carrier, Cheever, Rosen, Benitez, and Chang [1] who found that there were no significant multitasking behavior patterns as sign of digital generation across different age groups in the United States. Other surveys done in the U.K. [3, 11]; and Australia [2, 7] have also shown similar pattern. While reporting themselves as more engaged in using different kinds of technologies, those assumed as digital generation are too heterogeneous to be classified as a group: only a minority of them is truly technological savvy. Most of individual fall in this group are not more technologically illiterate than their previous generation.

However, results from this survey came from limited numbers of respondents. It seems that the digital generation of Indonesia behavioral patterns cannot be described with a limited number of respondents in the study. In addition, Indonesia also faced the inclusion of Internet technology. The responsiveness of inter-generational Internet users in Indonesia still needs to be studied further.

References

1. Carrier, L.M., Cheever, N.A., Rosen, L.D., Benitez, S., Chang, J.: Multitasking across generations: Multitasking choices and difficulty ratings in three generations of Americans. Computers in Human Behavior 25, 483–489 (2009)
2. Combes, B.: Generation Y: Are they really digital natives or more like digital refugees? Synergy 7(1), 31–40 (2009)
3. Helsper, E., Enyon, R.: Digital natives: Where is the evidence? British Educational Research Journal, 1–18 (2009)
4. Hill, D.T., Sen, K.: Wiring the warung to global gateways: The internet in Indonesia. Indonesia 63, 67–89 (1997)
5. Hill, D.T., Sen, K.: Internet in Indonesia's New Democracy. Routledge, London (2005)
6. Jones, C., Ramanau, R., Cross, S., Healing, G.: Net generation or Digital Natives: Is there a distinct new generation entering university? Computers & Education 54(3), 722–732 (2010)
7. Kennedy, G., Krause, K., Judd, T., Churchward, A., Gray, K.: First year students' experiences with technology: Are they really digital natives? Preliminary Reports of Findings, Biomedical Multimedia Unit of University of Melbourne Research (2006)
8. Kvavik, R.B.: Convenience, Communications, and Control: How Students Use Technology. In: Oblinger, D., Oblinger, J. (eds.) Journal of Educating Net Generation. Educase (2005), http://www.educause.edu/educatingthenetgen/
9. Leung, L.: Net-Generation Attributes and Seductive Properties of the Internet as Predictors of Online Activities and Internet Addiction. Cyberpsychology & Behavior 7(3), 333–348 (2004)
10. Lim, M.: Internet, Social Networks, and Reform in Indonesia. In: Couldry, N., Curran, J. (eds.) Contesting media power: Alternative media in a networked world. Rowman & Littlefield, Maryland (2003)

11. Margaryn, A., Littlejohn, A.: Are digital natives a myth or reality?: Students' use of technologies for learning. Computers & Education 56(2), 1–30 (2008)
12. Oblinger, D., Oblinger, J.: Journal of Educating Net Generation. Educase (2005), http://www.educause.edu/educatingthenetgen/
13. Palfrey, J., Gasser, U.: Born Digital: Understanding the First Generation of Digital Natives. Basic Books, New York (2008)
14. Prensky, M.: Digital Natives Digital Immigrants. On the Horizon 9(5) (2001)
15. Ramdhani, N.: Komunikasi berbasis Paperless Office: Studi tentang PLO Psikologi UGM. Jurnal Psikologi (dalam proses penerbitan) (2012)
16. Rosen, L.D.: Me, MySpace, and I: Parenting the Net Generation. Palgrave-MacMillan, New York (2007)
17. Tapscott, D.: Grown Up Digital: How the Net Generation is changing Your World. McGraw-Hill, New York (2009)
18. Wiradhany, W.: Uji Validitas dan Reliabilitas Alat Ukur Trust pada Pengguna E-Commerce di Indonesia. Skripsi, tidak diterbitkan. Universitas Katolik Indonesia Atma Jaya, Jakarta (2010)
19. Young, K.S.: Internet Addiction: The Emergence of a New Clinical Disorder. CyberPsychology and Behavior 1(3), 237–244 (1998)

End-to-End Delay Performance for VoIP on LTE System in Access Network

Ng Liang Shen[1,*], Noraniah Abdul Aziz[2], and Tutut Herawan[3]

[1] Faculty of Computer System and Software Engineering
Universiti Malaysia Pahang Lebuhraya Tun Razak, 26300 Gambang, Malaysia
[2] TATI University of College 24000 Kemaman Terengganu, Malaysia
[3] Department of Mathematics Education Universitas Ahmad Dahlan
Yogyakarta 55166, Indonesia
ngliangshen@gmail.com, noraniah@tatiuc.edu.my, tutut@uad.ac.id

Abstract. The desire Quality of Service (QoS) of Voice over Internet Protocol (VOIP) is of growing importance for research and study Long Term Evolution (LTE) is the last step towards the 4th generation of cellular networks. This revolution is necessitated by the unceasing increase in demand for high speed connection on LTE networks particularly for under variable mobility speed for VoIP in the LTE. This paper mainly focuses on performance of VOIP and the impact of resource limitations in the performance of Access Networks particularly important in regions where Internet resources are limited and the cost of improving these resources is prohibitive. By determine rate communication quality, is determined by end to end delay on the communication path, delay variation, packet loss. These performance indicators can be measured and the contribution in the access network can be estimated using simulation tool OPNET Modeler in varying mobility speed of the node. The overall performance of VOIP thus greatly improved significantly by deploying OPNET Modeler.

Keywords: Quality of Service (QoS), Voice over Internet Protocol (VOIP), Internet, performance, Future, Network Access, analysis, delay.

1 Introduction

The startling growth of Internet technology, coupled with the relatively low deployment cost of IP networks, has pushed for an integrated "IP-based core" - a single network for data, video and voice access. However, the diverse service-requirements and novel traffic characteristics of the emerging Internet applications have posed many technical challenges that the Internet community must address in the near future, as the emerging multimedia applications begin to constitute an ever-increasing fraction of Internet traffic. High quality interactive voice and video applications can tolerate little delay variation and packet loss. Quality of Service (QoS) is a defined level of performance in a data communications system. As an

* Corresponding author.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 162–172, 2013.

example, to ensure that real time voice and video are delivered without irritating blips, a guaranteed bandwidth is required. The plain old telephone system (POTS) has delivered the highest quality of service for years, because there is a dedicated channel between parties. However, when data is broken into packets that travel through the same routers in the LAN or WAN with all other data, QoS mechanisms are the only way to guarantee quality by giving real time data priority over non-real time data. A large number of factors are involved in making a high-quality VoIP call. These factors include the speech codec, packetization delay, packet loss, delay (coding, transmission, propagation and queuing), delay variation, and the network architecture to provide QoS. Other factors involved in making a successful VoIP call include the call setup signalling protocol, call admission control, security concerns, and the ability to traverse Network Access Translation (NAT). Although VoIP involves the transmission of digitized voice in packets, the telephone itself may be analog or digital. The voice may be digitized and encoded either before or concurrently with packetization. In making a high quality VOIP call and the engineering trade-offs that must be made between delay and the efficient use of bandwidth [1,3]. VoIP becoming popular can be mainly attributed to the cost advantages to consumers over traditional telephone networks. The traditional business model for telephone services has been that most people pay a flat monthly fee for local telephone call service and a per-minute charge for long-distance calls.Introduction of Long Term Evolution (LTE), the 4th Generation (4G) network technology release 8 specifications are being finalized in 3GPP have developed and planning to globalize extensively compared to 3rdGeneration (3G) and 2nd Generation (2G) networks [1]. LTE determines goals peak data rate for Downlink (DL) 100 Mbps and Uplink (UL) data rate for 50Mbps, increased cell edge user throughput, improved spectral efficiency and scalable bandwidth 1.4 MHz to 20 MHz [2]. VoIP capacity of LTE has to show better performance as Circuit Switch voice of UMTS. LTE should be at least as good as the High Speed Packet Access (HSPA) evolution track also in voice traffic. It discusses the challenging issues that need to be faced by computer networks to transmit the VoIP applications. It gives the description idea about the VoIP over LTE and their functionality and design parameters of the LTE networks. After determining the problems it is necessary to identify the research questions that lead the research process to be in the scope [2].

a) How much the maximum throughput is support in the different bandwidth (e.g. 1.4 MHz, 3 MHz, 5 MHz, 10 MHz, 15 MHz and 20 MHz)?
b) What is the impact on the VoIP quality in terms of E2E delay when the network is congested with VoIP only?
c) To what extent do the performances of packet loss for interactive voice vary?

The 3GPP LTE is a new standard with comprehensive performance targets, therefore it is necessary to evaluate the performance and stability of this new system at an early stage to promote its smooth and cost-efficient introduction and deployment. The motivation behind the design models presented in this report is to discuss issues related to traffic behavior for VoIP alone as well as along with other traffic in the LTE network. E2E delay for VoIP is a matter of fact for performing real-time application efficiently over the Internet. Today, emergence of the real-time application demands more resources. The main motivation of our paper work is to ensure fast and reliable

voice communication for huge number of users in wireless network. LTE evolved packet system (EPS) is the bearer of the QoS level of granularity. This system also establishes the packet flow between the user terminal (UE or MS) and the packet data network gateway (PDN-GW). The traffic running between a particular client application and the service can be wrecked into split service data flows (SDFs). Mapping the same bearer, SDFs receive common QoS activities (e.g., scheduling policy, queue management policy, rate shaping policy, and radio link control (RLC) configuration). A scalar value referred to as a QoS class identifier (QCI) with the help of bearer, specifies the class to which the bearer belongs. Set of packet forwarding treatments referred by QCI (e.g., weights scheduling, ad-mission thresholds, configuration of link layer protocol and queue management thresholds) preconfigured through the operator on behalf of each network element. The class-based technique applies in the LTE system to improve the scalability of the QoS framework. In the LTE framework, bearer management and control follows the network-initiated QoS control paradigm that initiated network establishment, modification, and deletion of the bearers. Two types of bearers in LTE [2,3,4]:

a. **Guaranteed bit rate (GBR):** Dedicated network resources correlated to a GBR value connected with the bearer and permanently allocated when a bearer becomes established or modified.
b. **Non-guaranteed bit rate (non-GBR):** In the LTE system, non-GBR bearer is as-signed as the default bearer, similar to the preliminary SF in WiMAX, used to establish the IP connectivity. A non-GBR bearer has enough knowledge about congestion-related packet loss. In the framework, additional bearer is assigned as a dedicated bearer which is GBR or non-GBR.

Future Research need to cover further analyses could be carried out using the measurement data captured. A synthetic workload model for VoIP could be further developed, based on the workload characteristics identified in the measurement study.

2 How Is the Performance for VoIP on LTE System in Access Network

In LTE, the core network operations are completely based on packet switching domain, example, all the network interfaces are dependent on IP protocols, and hence it is known as Evolved Packet Core (EPC). The essence of EPC is to keep the number of operating nodes and interfaces as minimum as possible. The EPC divides the network components into control-plane objects such as data/barer-plane entity (i.e. a Serving Gateway) and the Mobility Management Entity (MME). The MME is considered as a signaling entity and used to represent the control plane function of the EPC. Such control functions include, among others, location function, the subscribers' equipment paging, and the bearer establishment and the connections establishment, roaming management [4]. The fundamental architecture of LTE system is presented in Figure 1. All the network interfaces are based on internet protocols (IP). The LTE system comprised of the core network and radio access network which represent the IP connectivity layer of LTE system.

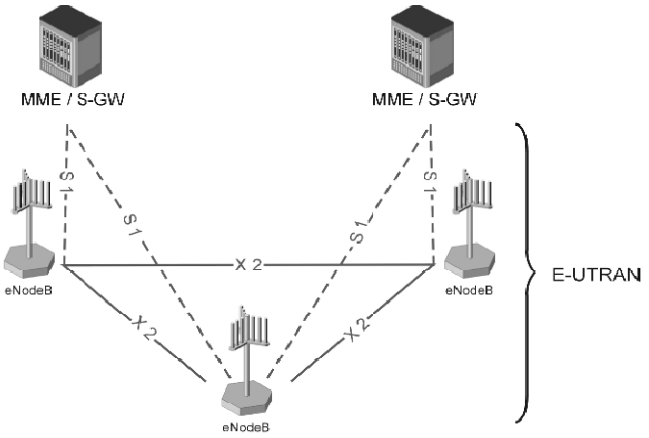


Fig. 1. Networks Architecture of LTE [16]

The Serving Gateway (S-GW) functions as switching as well as routing node to route and forward the data packets to and from the Base or Evolved-Universal Terrestrial Radio Access Network NodeB (eNB) [5]. Between the EPC and the external packet data network, a PDN-GW is often used as an inter-face point or an edge router. It is also possible that a UE has synchronized connectivity with more than one PDN GW. The responsibilities of the PDN-GW include establishment, maintenance, and deletion of GTP tunnels to S-GW or SGSN in the case of inter-RAT mobility scenarios. The PDN-GW routes the user plane packets by allocating the user's dynamic IP addresses. Apart from that, it provides functions for lawful interception, policy/QoS control, and charging [6].

3 LTE QoS Framework and Voice over IP (VoIP)

LTE evolved packet system (EPS) is the bearer of the QoS level of granularity. This system also establishes the packet flow between the user terminal (UE or MS) and the packet data network gateway (PDN-GW). The traffic running between a particular client application and the service can be wrecked into split service data flows (SDFs). Mapping the same bearer, SDFs receive common QoS activities (e.g., scheduling policy, queue management policy, rate shaping policy, and radio link control (RLC) configuration) [19, 20]. A scalar value referred to as a QoS class identifier (QCI) with the help of bearer, specifies the class to which the bearer belongs. Set of packet forwarding treatments referred by QCI (example, weights scheduling, admission thresholds, configuration of link layer protocol and queue management thresholds) preconfigured through the operator on behalf of each network element [7]. The class-based technique applies in the LTE system to improve the scalability of the QoS framework. [8].

3.1 Two Types of Bearers in LTE

There are two types of bearer in LTE which are GBR and non-GBR. In the framework, additional bearer is assigned as a dedicated bearer which is GBR or non-GBR. Real Transport Protocol (RTP) is the basic protocol in VoIP engineering, which is, not only for transporting media streams but also to initialize the media session in concord with SIP. It is also used for media stream supervision and intended to provide out-of-band control information for the RTP flow. In response to the media quality that supplies to the other members in the media session via separate UDP port, there are many additional functionalities of RTP. Audio and video synchronization and quality improvements through low compression instead of high compression are a few of them. VoIP transmissions are deployed through traditional routing. A typical VoIP structural design is shown on the Figure 2, though many "possible" modifications of this architecture are implemented in existing systems. In VoIP engineering, original voice signal is sampled and is encoded to a constant bit rate digital stream at the end of the sending process.In place of circuit-switched voice transmission and traditional dedicated lines, these packets flow over a general-purpose packet-switched digital to analog signal in the receiving end for it to be easily detected. The VoIP infrastructure can be visualized as three layers: end user equipment, network components, and a gateway to the traditional telephone network as stated in Figure 3.

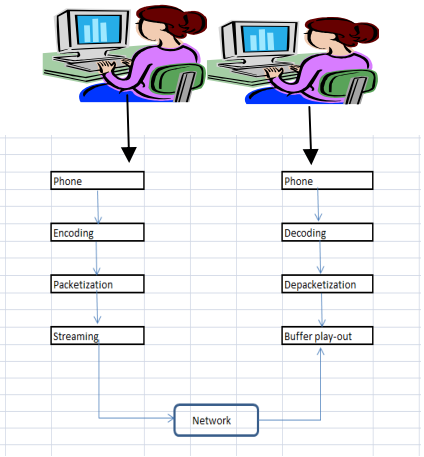


Fig. 2. VoIP Architecture of LTE

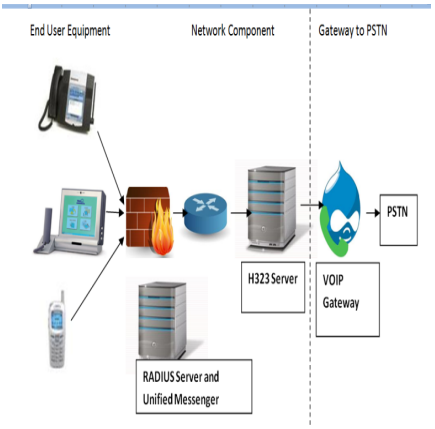


Fig. 3. VOIP. Network Archtecture

3.2 Characteristics VoIP Traffic Delay

AMR codec provides the VoIP traffic along with the Voice Activity Detector, Relieve Noise Generation and Discontinuous Transmission. Depending on the speed activity of the traffic, AMR provides a constant rate of small packets transmission. During the active period, one VoIP packet took at 20 ms intervals and 160 ms interval for one Silence Description (SID) packet during silent period. To improve the spectral

efficiency of the VoIP traffic, UDP, IP and RTP headers in LTE are also compressed with Robust Header Compression (ROHC). According to [9], for voice signal, 250 ms is the maximum tolerable mouth-to-ear delay and around 100 ms delay for the Core Network and also less than 150 ms acceptable delay for Medium Access Control (MAC) buffering and Radio Link Control (RLC). Both end users are LTE users and assume less than 80 ms acceptable delay for buffering and scheduling. For 3 GPP performance evaluations 50 ms delay has been bound for variability in network end-to-end delays. End-to-end delay means the time required for a packet to be traversed from source to destination in the network and is measured in seconds. Generally, in VoIP network there are three types of delays occurring during the packet transverse. They are: sender delays when packets are transverse from source node, network delay and receiver delay. In one direction from sender to receiver for VoIP stream flow, end-to-end delay can be calculated by the equation. Let say in seconds-that a network needs to send a message with n bytes as $T(n) = \alpha + \beta n$ In mathematical terms, this is a line equation, where α is the *constant*, and β is the line *slope* or *gradient* [9,10].

4 Simulation Design and Methodology

In real world scenarios, performance evaluation of a well-designed network model and the model itself carries significant importance. To cope with the challenge, practice using simulator OPNET (Optimized Network Engineering Tool), as kernel source code is not open source. [11,12]. In order to determine the maximum number of calls that can be supported by an existing network, while limiting VoIP delay constraint utilize queuing analysis to approximate and determine the maximum number of calls that the existing network can support while maintaining a delay of less than 80ms.,the Principles of Jackson theorem for analysing queing [12]. Analysis by decomposition is summarized in first isolating the queuing network into subsystems, example, and single queuing node[13,14,15,16]. Next, analysing each subsystem separately, considering its own network surroundings of arrivals and departures. Then, finding the average delay for each individual queuing subsystem finally, aggregating all the delays of queuing subsystems to find the average total end-to-end network delay [17,18,19,20].

Figure 4 presents an algorithm computes the maximum number of calls considering VoIP delay constraint.

a. Initially, no calls are introduced only traffic in the network background traffic.
b. A new call is added, according to the call distribution.
c. For each network element, $\lambda = \lambda VoIP + \lambda bg$ is computed. Where, $\lambda VoIP$ is the total added new traffic from a single VoIP in pps, and λbg is the background traffic in pps). The value λbg is known for each element; however, $\lambda VoIP$ can get affected by introducing a new call depending on the call traffic flow, i.e. whether or not the new call flow passes through the network element.
d. For each network element, the average delay of a VoIP packet is computed.
e. The end-to-end delay is computed by summing up all the delays of step (d) encountered for each possible VoIP flow. This includes all external and internal flows, with internal flows consisting of intra-floor and inter-floor.

f. If the maximum network delay is less than 80 ms, then the maximum number of calls has not been reached. Therefore a new call can be added, and hence go to steps (b)-(f). If not, the maximum delay has been reached. Therefore the number of VoIP calls bounded by the delay is one less than the last call addition.

Input n : number of network of elements
$\lambda[1..n]$: background traffic for network element 1,2, … n
Delay [1.. n]: delay for network element 1,2 … n
P: set of call − flow paths (p)where p is a subset of {1,2, … n}
Output: MaxCalls: maximum number of calls
$\lambda Voip \leftarrow 100pps$, or 180.8kbps;
VoIP MaxDelay $\leftarrow 80$; // network delay for VoIP call in ms
MaxDelay $\leftarrow 0$;
MaxCalls $\leftarrow -1$;
Delay [1.. n] $\leftarrow 0$;
While MaxDelay $< VoIP\ MaxDelay\ Do$
MaxCalls $\leftarrow$ MaxCall + 1
Generate a call according to call distribution and let P_C be its flow path
for each element i in Pc do
$\lambda i \leftarrow \lambda i +$, or $\lambda voip$
if i is a link then
Delay $\leftarrow (1 - \lambda 2\mu i)/(\mu i - \lambda i)$
Else
Delay $\leftarrow 1/(\mu i - \lambda i)$
end if
end for
for each p in P where p ∩ $P_C \neq \emptyset$ do
PathDelay (p) $\leftarrow \sum_{n=1}^{\infty}$ Delay where i is a network element in path p
if PathDelay (p) $> MaxDelay\ then$
MaxDelay ← PathDelay (p)
end if
end for

Fig. 4. Algorithms number of calls VoIP [8]

5 Result and Analysis

5.1 End-to-End (E2E) Delay Performance

For VoIP applications, the packet E2E delay should not exceed 150 ms to evaluate that the quality of the created VoIP calls are accepted. In this section, the packet E2E delays result for different scenarios are presented in various statistical plots. Scenario 1, correspond to the Baseline VoIP, while Scenario 2 correspond to the Congested with VoIP. In all the scenarios, the sources and destinations happen to be started at 100 seconds as the start time of the profile and application configuration has been set to 40 seconds and 60 seconds, respectively. the figures presented in following sections, X axis represents simulation time in seconds; Y axis represents in seconds. Figure 5 illustrates the comparable performance of the E2E delay under the different case scenarios of Baseline VoIP network. E2E delay is measured for the VoIP traffic

flows between source node and destination node through Baseline network with various node speeds the blue line shows the E2E delay of the scenario where the node speed is fixed (0 mps). The blue line show the E2E delay of scenarios where the node speeds are 23.83 mps, whereas red line shows in Figure 6, the congested VoIP network level of about 60.84 mps. With the increasing simulation time, the both curves settle at Simulation time 2.5 sec.

Table 1. Statistics of E2E delay of Based-line Congested VoIP Network

Baseline Before Congestion	Congestion End to End Delay	Simulation time	Baseline Before Congestion	Congestion End to End Delay	Simulation time
23	75	3.26	48.82	49.18	1.01
23.83	74.17	3.11	49.66	48.34	0.97
24.67	73.33	2.97	50.49	47.51	0.94
25.50	72.50	2.84	51.32	46.68	0.91
26.33	71.67	2.72	52.16	45.85	0.88
27.17	70.84	2.61	52.99	45.01	0.85
28.00	70.00	2.50	53.82	44.18	0.82
28.83	69.17	2.40	54.65	43.35	0.79
29.66	68.34	2.30	55.49	42.51	0.77
30.50	67.50	2.21	56.32	41.68	0.74
31.33	66.67	2.13	57.15	40.85	0.71
32.16	65.84	2.05	57.99	40.01	0.69
33.00	65.00	1.97	58.82	39.18	0.67
33.83	64.17	1.90	59.65	38.35	0.64
33.83	64.17	1.90	60.49	37.52	0.62
34.66	63.34	1.83	61.32	36.68	0.60
35.50	62.51	1.76	62.15	35.85	0.58
36.33	61.67	1.70	62.98	35.02	0.56
37.16	60.84	1.64	63.82	34.18	0.54
37.99	60.01	1.58	64.65	33.35	0.52
38.83	59.17	1.52	65.48	32.52	0.50
39.66	58.34	1.47	66.32	31.68	0.48
40.49	57.51	1.42	67.15	30.85	0.46
41.33	56.67	1.37	67.98	30.02	0.44
42.16	55.84	1.32	68.82	29.19	0.42
42.99	55.01	1.28	69.65	28.35	0.41
43.83	54.18	1.24	70.48	27.52	0.39
44.66	53.34	1.19	71.31	26.69	0.37
45.49	52.51	1.15	72.15	25.85	0.36
46.32	51.68	1.12	72.98	25.02	0.34
47.16	50.84	1.08	73.81	24.19	0.33
47.99	50.01	1.04	74.65	23.35	0.31

Packet loss refers to the failure of one or more transmitted packets to reach their destination across a network. A VoIP user is satisfied if more than 98 % of its voice packets are delivered successfully. Packet loss is determined using the following equation.

$$\left(\text{Packet Loss} = \frac{Send\ Packet - Received\ Packet}{Send\ Packet} \times 100\right)$$

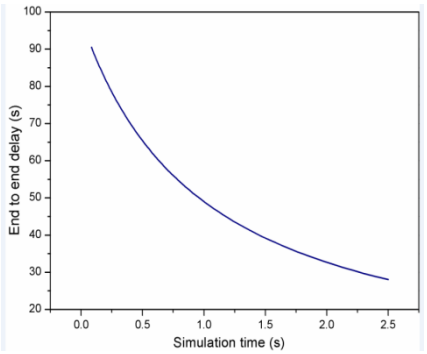


Fig. 5. End-to-End Delay of Baseline Network

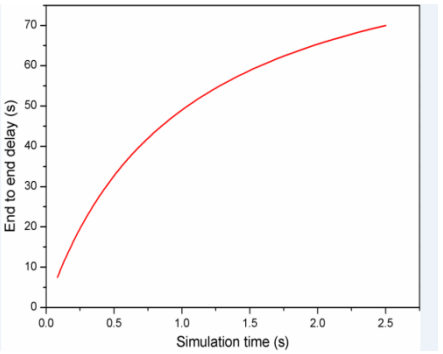


Fig. 6. End-to-End Delay of Congested VoIP Network

5.2 Packet Loss Performance for Baseline VoIP Network

In congested VoIP network. Can observe in figure, sent and received data are overlapped between 100 seconds to 140 seconds in all the simulation scenarios. After 140 seconds when the load is around 95%, the packet loss is found in all the cases. In the Congested VoIP and VoIP congested with FTP network scenario, on an average, packet loss in the VoIP Congested with FTP network for case 1 is about 74% higher than that of Congested VoIP network. In other both cases, average packet losses are 88%, 95% and 84% higher respectively. To wrap it, all of the voice traffic corresponding to each case in VoIP Congested with FTP experience higher packet loss than Congested VoIP network.

6 Conclusion

Comprehensive study, analysis and evaluation of the end-to-end delay performance evaluation for VoIP in the LTE network have been done. The evaluation is made by simulating in OPNET Modeler based on performance metrics such as end-to-end delay and throughput. Three network scenarios have been simulated: Baseline VoIP network scenario, congested VoIP network scenario and VoIP congested with FTP network scenario. It has been found that maximum throughput increased as the bandwidth incremented. Out of the three methods simulated for throughput measurement, the scenario with highest bandwidth (example 20 MHz) exhibited maximum throughput. After that, quality measurement of VoIP has been done with respect to E2E delay, both for a network congested exclusively with VoIP and VoIP with FTP. Four scenarios have been created for this evaluation, one case with stationary node and other three cases with mobile nodes (gradually increasing the node speed). The simulation results

showed that when the node is not moving, E2E delay is slightly higher for network congested with VoIP only. In other cases, better E2E is obtained for this network due to the presence of moving node. Finally the rate of packet loss for the remains quite minimal for congested with VoIP network regardless of the node speed. Mean-while, for VoIP with FTP network, packet loss rate is also quite insignificant for fixed node case but the rate upshots as the node starts moving [19,20].

References

1. 3GPP, Release 8 V0.0.3.: Overview of 3GPP Release 8: Summery of all Release 8 Features (November 2008)
2. Puttonen, J., Puupponen, H.-H., Aho, K., Henttonen, T., Moisio, M.: Impact of Control Channel Limitations on the LTE VoIP Capacity. In: The Proceedings of 2010 Ninth International Conference on Networks, pp. 77–82 (2010)
3. Puttonen, J., Henttonen, T., Kolehmainen, N., Aschan, K., Moisio, M., Kela, P.: Voice-Over-IP Performance in UTRA Long Term Evolution Downlink. In: The Proceedings of IEEE Vehicular Technology Conference, pp. 2502–2506 (2008)
4. Khan, F.: LTE for 4G Mobile Broadband: Air Interface Technologies and Performance, 1st edn. University Press, Cambridge (2009)
5. Martín-Sacristán, D., Monserrat, J.F., Cabrejas-Peñuelas, J., Calabuig, D., Garrigas, S., Cardona, N.: On the Way towards Fourth-Generation Mobile: 3GPP LTE and LTE-Advanced. EURASIP Journal on Wireless Communications and Networking, 1–10 (2009)
6. Holma, H., Toskala, A.: LTE for UMTS - OFDMA and SC-FDMA Based Radio Access. John Wiley & Sons Ltd., Finland (2009)
7. Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestri-al Radio Access Network (E-UTRAN). 3GPP TS 36.300 version 8.11.0 (Janauary 2010)
8. Olorunda, O., Olorunda, A.: Bridging the Digital Divide - The Social and Cultural Impact of VoIP in Developing Countries: Nigeria as a Case Study. In: The Proceedings of PTC 2006 (2006)
9. One Way Transmission Time. ITU-T recommendation G.114 (May 2003)
10. M. J. Karam and F. A. Tobagi.: Analysis of delay and delay jitter of voice traffic in the Internet. Speedtrak Communications, San Mateo, CA 94401, USA, Computer Systems Laboratory, Stanford University, Vol. 40, pp. 711-726 (2002)
11. OPNET. Opnet modeler 16.0
12. `http://www.opnet.com/solutions/network_rd/modeler.html`
13. Salah, K., Alkhoraidly, A.: An OPNET-based simulation approach for deploying VoIP. International Journal of Network Management 16(3), 159–183 (2006)
14. Dantu, R., Fahmy, S., Schulzrinne, H., Cangussu, J.: Issues and challenges in securing VoIP. Computers & Security 28, 743–753 (2009)
15. Salah, K., Almashari, M.: An Analytical Tool to Assess Readiness of Existing Networks for Deploying IP Telephony. In: The Proceeding of 11th IEEE Symposium on Computers and Communications, ISCC 2006, pp. 301–305 (2006)
16. Masum, M.E., Babu, M.J.: End-to-End Delay Performance Evaluation for VoIP in the LTE network. Postgraduate Thesis, Blekinge Institute of Technology 23 (2011)

17. Gupta, J.D.: Performance issues for VOIP in Access Networks. Postgraduate Thesis, Department of Mathematics and Computing Faculty of Sciences, University of Southern Queensland, 87–96 (2005)
18. He, Q.A.: Analysing the Characteristics of VoIP Traffic. Postgraduate Thesis, Department of Computer ScienceUniversity of Saskatchewan, 62–77 (2007)
19. Muhleisen, M., Walke, B.: Evaluation and Improvement of VoIP Capacity for LTE. ComNets Research Group, pp. 1–7. Aachen University, Germany
20. Holma, H., Toskala, A.: LTE for UMTS: OF DMA and SC-FDM Based Radio Access, pp. 259–280. John Wiley & Sons (2009)

Mobile Collaboration Technology in Engineering Asset Maintenance – What Technology, Organisation and People Approaches Are Required?

Faisal Syafar and Jing Gao

School of Information Technology and Mathematical Sciences,
University of South Australia, Australia
Faisal@mymail.unisa.edu.au, Jing.Gao@unisa.edu.au

Abstract. Engineering asset maintenance consists of coordinated activities and practices for retaining or restoring a piece of equipment, machine, or system to specified operable conditions to achieve its maximum useful life. An integrated high-level maintenance comprising multiple sub-systems requires the collaboration of many stakeholders including multiple systems and departments. Several of specialised technical, operational and administrative systems have been invested by engineering asset organisations to enhancing their asset management and maintenance systems, however there is no common ground among engineering asset organisations about what are collaborative maintenance are required for adoption/implementation. The lack of systematic approach, together with the lack of specific requirements to implement mobile collaborative maintenance requests a comprehensive framework for guiding engineering organisation to implement of new mobile technologies that meet all maintenance collaboration requirements. This research proposes to develop an appropriate mobile collaboration framework based on Delphi and Case Study investigation.

Keywords: Mobile technology, Collaboration, Engineering asset, Framework.

1 Introduction

An imperative element of business management is having engineering asset management (EAM). It is essential to realize a business organization's mission which is having or operate assets. EAM is concerned with the life cycle of engineering or physical assets. It is also a critical stage of the continuous life cycle of assets, which includes acquisition, design, installation, maintenance and operation, and disposal stages [1]. The purpose of EAM is to maximise the total benefits of an enterprise by effectively using these assets throughout the whole life cycle.

The importance of maintenance function has increased because it plays an important role in retaining and improving system availability and safety, and product quality [2]. Reference [3] states that engineering assets in industries rely highly on their maintenance division to maintain and ensure assets are delivered properly. This

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 173–182, 2013.

author also revealed that in the last 30 years, the practice of doing maintenance has significantly changed due to developments in equipment design, information and communication technology, cost pressures, customer acceptance of risk and failures [3] and the existence of multiple stakeholders and departments [4]. Moreover, current working circumstances are more complex and therefore need to be managed by multiple and interlinked activities [5]. Hence, an integrated high-level maintenance system which contains multiple sub-systems requires the collaboration of multiple stakeholders such as departments or units to improve resources, information sharing and maintenance practices.

Collaborative maintenance is not a technology or a software solution; rather, it is a customized business strategy unique to each situation [6]. Based on a review of some relevant references [7][8][9], it is found that many organizations already have a collaborative maintenance system in place. However, with proper collaboration and commitment, that system can be expanded in scope and effectiveness.

To achieve the quality and efficiency of maintenance for engineering assets, this research proposes a framework to guide the adoption and implementation of new mobile technologies. It aims to facilitate asset maintenance collaboration, where organizations can expand their existing technology. In this context, efficiency means maintaining engineering assets without interrupting the production process for unnecessary breakdowns.

This research is structured as follows: the first section describes the motivation of this research. The second section provides a brief review of the concept of mobile collaboration technology for engineering or physical asset maintenance (PAM). This is followed by a brief description of the proposed framework, research methods, and the last section outlines a conclusion.

2 Research Motivation

This research is motivated by five factors. **First**, today's asset maintenance practices rely on access to information and team expertise from dispersed sites. Many businesses or companies have several interdependent departments and sub-systems that collaborate on various issues. Maintenance personnel in the form of individual and/or groups communicate, coordinate, integrate and distribute work. **Secondly**, information exchange in the form of direct communication, discussion, negotiation and decision making to complete the integrated maintenance (including strategic, tactical and operational levels) as well as maintenance information systems for structures a share pool of maintenance knowledge between certain maintenance roles (managers, directors, supervisors, engineers, technicians) in different site (between maintenance personnel in different offices, between maintenance and a remote help desk expertize centre, and between maintenance personnel in the field force in different sites) in any time are not supported by current PAM system/technology. **Third**, the emerging trend of mobile technologies is rapidly developing and they are viewed as business enablers, and have the potential to support asset maintenance practice. **Fourth**, some frameworks for mobile collaboration have been identified in this preliminary literature review, however, only a few of these frameworks are relevant and applicable in

this research. Most of the design frameworks in the last decade refer to the technological approaches for hardware, software and network. **Last but not least**, only a few studies have explored how mobiles in the context of technological, organizational, and personnel requirements support collaborative maintenance in engineering asset organizations. Few survey and case studies have been conducted to understand the current state of collaboration technologies including mobile technology support engineering asset organizations.

3 Literature Review

3.1 Engineering Asset Maintenance

Maintenance is a combination of actions intended to retain an item in, or restore it to, a state in which it can perform the function that is required for the item to provide a given service. This concept leads to a first classification of the maintenance actions in two main types: actions oriented towards retaining certain operating conditions of an item and actions dedicated to restoring the item to supposed conditions. “Retention” and “restoration” are denominations for action types that are then converted into “Preventive” and “Corrective” maintenance types in the maintenance terminology by European Committee for Standardization.

Reference [10] acknowledges that maintenance processes consists of several tasks that must be in line with three levels of business activities, namely: strategic, tactical and operational. At the *strategic* level, business priorities will be transformed into maintenance priorities. The transformation process is done by supporting expertise within a certain period to address current and/or potential gaps in equipment maintenance action or strategy. A generic maintenance plans (middle and long range) will be obtained at this level. Actions at the *tactical* level will define the appropriate task of maintenance resources such as skills, materials, test equipment, etc. to fulfil a business’s maintenance plan. At this level, a detailed program consists of particular tasks and the resources allocated would be achieved. At the *operational* level, maintenance tasks would be done accurately by skilled engineers, in the time that has been allocated, following right procedures, and using the appropriate tools. Preventive and corrective maintenance procedures are carried out so tasks receive a high level of attention. All maintenance tasks will be completed and the maintenance data history would be recorded in the information system at this level.

3.2 Current IT/IS Capabilities for Engineering Asset Maintenance

Reference [11] emphasise that in order to manage the sophisticated AM process and to provide its data requirements, particular technology and systems are required. The system that captures, maintains, and manages all the needed asset information throughout the entire asset lifecycle is critical in providing effective asset management. Currently, several of specialised technical and operational systems have been invested by EAM organisations to enhancing their asset maintenance systems. These technologies and systems aimed at support the whole asset lifecycle.

The very popular maintenance information systems that have been implementing for engineering asset maintenance are Computerised Maintenance Management Systems (CMMS) [12]. However, although CMMS makes a great volume of information available for reliability and efficiency analysis of the delivery of the maintenance function, most experts agree that successful CMMS is less than 30% of total CMMS applications [13]. The main reasons are [14]: Selection errors , insufficient commitment, lack of training , failure to address organizational implications, underestimating the project task, lack of project resources and lack of demonstrable use of system output.

3.3 Collaboration Requirements by Asset Maintenance' Stakeholders

Mobile technologies play a key role in this setting, facilitating to establish tightly integrated environments between different groups and organizations that bear stakes on the performance of the industrial assets [15]. Despite the fact that the use of advanced application solutions in manufacturing, production, or process facilities takes place at a different scale, the emerging trend has already shown that mobile technologies have a great potential to redefine and re-engineer the conventional setting. They have already begun to offer advanced and smart solutions to remotely manage complex, high-risk, and capital-intensive assets, regardless of the geographical location, building agile information and knowledge networks [16].

In order to encounter good asset maintenance and meet the optimum performance of engineering assets, organisations require a collaborative teamwork within key functional areas (stakeholders) of the engineering organisations. Shared understanding, coordination, cooperation and collaboration across maintenance stakeholders of what asset maintenance is and how the entire maintenance team influence the ability to achieve organisational objectives through those assets are one of the critical success factors of asset management. Collaborative asset maintenance is applicable to all those who have a role in the maintenance of engineering assets including directors, managers, supervisor, engineers, IT and maintenance technicians.

Mobile collaboration technology required for asset maintenance need to be capable of simultaneously handling, processing and delivering technical and operational information to multiple maintenance crew at multiple locations at any time to enhance asset maintenance planning and implementation within the three levels of business activities. The requirements are including technological, organisational, as well as personal perspectives.

3.4 Proposed Framework

In order to develop the mobile collaboration PAM framework, the research questions need to be answered. The major question is: How can mobile collaboration technologies assist asset maintenance in engineering asset management organization?

Based on the extensive literature review, a conceptual research framework was developed as shown in Figure 1. It encapsulates the core concept of [17] TOP model as a means of studying collaboration requirements from either technical or organisational or personal perspectives. It also includes the alignment of maintenance processes with three levels of business activities: strategic, tactical and operational [10]. This conceptual framework will guide the planning and activities in the subsequent research for investigating collaboration requirements in physical asset maintenance.

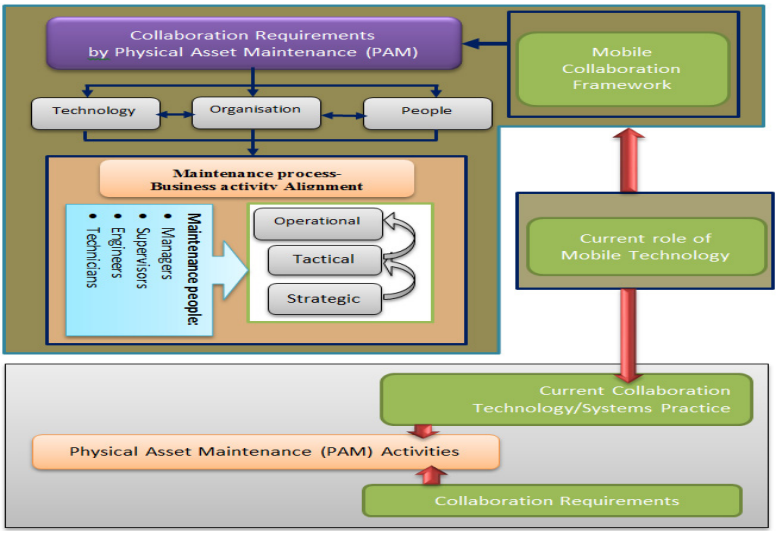


Fig. 1. Preliminary Maintenance Collaboration Framework

4 Research Methods

This research will be an interpretive study using both quantitative and qualitative methodologies. References [18][19][20] have reasoned that interpretive attempts to understand phenomena through the meanings that people assign to them are relevant. This understanding is particularly relevant in this research because the researcher is seeking to understand certain issues by industries survey, Delphi study and interviewing people on how mobile collaboration technologies will assist the asset maintenance process in a given organization's context. In order to create a complete set of requirements of collaboration maintenance in engineering organizations, the case study results here will be triangulated with the survey, Delphi study and case study findings. Triangulation is the use of more than one research strategy to explore the same phenomenon so that the credibility of research results is improved [21].

4.1 The Delphi Technique

This study is conducted to identify collaboration requirements, current collaborative maintenance practice and mobile technology roles in support collaborative engineering asset maintenance. The Delphi technique is employed to more accurately build the consensus from the panel expert's perception [22]. The Delphi study is a group process to solicit expert responses toward reaching consensus on a particular problem, topic, or issue by subjecting them to a series of in-depth questionnaires, interspersed with controlled feedback [22].

The Delphi method is employed for several reasons. The topic 'Mobile collaboration technology in engineering asset maintenance' is quite new, it is complex, a few literatures series have been found, and not much empirical data was available. Those are the reasons why Delphi study is useful to confront a mobile maintenance expert's panel. Delphi study is carried out in this research which comprised three rounds [23].

Nomination of Experts. A total of 47 experts who have strong academic backgrounds, research experience and professional in the area of mobile asset maintenance were invited to participate in the Delphi survey. Of these, 20 are willing to participate in the research project. They are 8 from universities and 12 professionals worldwide. The expert's profiles are illustrated below in Table 1 and 2.

Table 1. Participants by Role

Background of expert	Participants	
	Frequency	Percentage (%)
Academia	8	40
Professional	12	60
Total	20	100

Table 2. Participants by Geographic Location

Location of expert	Participants	
	Frequency	Percentage (%)
Australia	2	10
Canada	3	15
France	1	5
Germany	1	5
Greece	1	5
Malawi	1	5
Qatar	1	5
Singapore	1	5
United Arab Emirates	1	5
US	8	40
Total	20	100

Delphi Design. Three-round Delphi email-based questionnaire is designed. **The first round** is initial collection of requirements consisted of open-ended solicitation of ideas. Respondents were asked mainly about three basic questions, each corresponding to one of the research questions. The questionnaire asked experts to list general and the collaborative asset maintenance specific requirements, selecting criteria, benefits as well as initiatives issue that my hinder maintenance collaboration. **The second round** is validation categorized list of requirements. The experts were asked to verify the list that the researcher have correctly interpreted and placed them in an appropriate category/group based upon first round responses. In this round the experts were also requested to remove, added or regrouped the item (s) into other group/category. **The third round** is ranking relevant requirements. The consensus in the ranking order of the relevant group/category about requirements will be achieved in this final iteration. They will also be asked about the correlation between requirements (if any) as well as the critical requirements that need to be focus on.

4.2 Multiple Case Studies

Semi-structured interview-based multiple case studies will be conducted to explore the collaboration requirements for asset maintenance practices, to obtain information on the deficiencies in existing collaboration requirements.

In order to create a complete set of requirements of collaboration maintenance in engineering organizations, the case study results here will be triangulated with the Delphi study findings. Triangulation is the use of more than one research strategy to explore the same phenomenon so that the credibility of research results is improved [21].

5 Preliminary Findings and Discussion

Delphi question 1: ***Please mention collaborative maintenance requirements.***
From the responses of 19 panel members, we analyzed 63 individual statements. We then grouped into similar requirements and then mapped into Technology (T), Organization (O) and People (P) approaches as illustrated in Table 3.

Table 3. Collaborative maintenance Requirements

TOP	Delphi Round 1
Technology	*1. Synchronised multi-user access over a feature-populated dashboard*
	2. Data and services access through contextualised and mobile interfaces
	3. Data and services functionality porting to the cloud
	4. Autonomous information/communication exchange
	5. Linking the maintenance planning and dispatching
	6. Provides different mode for specific maintenance role
	7. Support interoperability between maintenance role
	8. Provide a platform for maintenance knowledge sharing across maintenance crew
	9. Social networking
Organization	*1. Cross-organisational management communication*
	2. Appropriate coordination mechanism of the team
	3. Availability and readiness all of maintenance crews
	4. Maintenance must be profit and customer-centred
	5. Clear maintenance vision (maintenance strategy-business objective)
	6. Maintenance crew using common maintenance language on syntactic and semantic consideration
	7. Combine professional experiences to support team work
	8. Provide team building activities to develop team work and skills

Table 3. (*continued*)

People	1. *Informal social networking between personnel*
	2. *Craft skill and training*
	3. *Common understanding of maintenance processes*
	4. *Common understanding of the system*

Delphi question 2: ***Please mention the current role of mobile technology in support asset maintenance collaboration technical/system***

We coded 19 responses into 42 individual statements. The statements were then clustered by similarity into categories and finally mapped to high-level feature areas as can be seen in Table 4.

Table 4. Mobile technology roles

Area	Feature category
Flexibility (initiate application at flexible sites in unstructured networked)	Visualising of collected data, parameter history and trending.
	Contextualising access over remote data and services.
	Critical for response time for data or information that can lead to early correction and or identification of failures.
	Providing the notification of failure through mobile devices
	Detecting the location of skilled maintenance personel nearby an asset that has experienced a failure through GPS.
	Mobile technology allows at the right place to access directly to a set of information coming from all the potential actors involved in the decision (CMMS, ERP, sensors, etc.).
Empowering management	Resources management (material, maintenance people) facilitator for continous task monitoring/assignment/reporting.
	Building and identifying process verification tasks, approvals.
	It helps to report failure effectively and report labors actual working hours and availability.
	Allowing to take the right maintenance decision, at the right time, at the right place, from the right information.
	Enhancing accuracy of critical data entry for maintenance history.
	Off-site (not in office) notifications and live feeds.
	Q/A decisions
Others	early adopters stage in the technology livecycle. stage.
	Extremely limited use at the moment

6 Conclusion

Through the development of mobile technologies, the processing of information can be performed by technical personnel away from the central production office or site. Maintenance personnel, when doing their tasks, require relevant information in different sites and need to communicate interactively with experts in the back office. Using mobiles allows maintenance personnel to continuously receive a daily schedule from the head office. This leads to the saving of time and improving customer service and profitability. Furthermore, it is expected that the research finding will develop a unique framework that addresses the following issues (1) Business process alignment at all three levels (strategic, tactical and operational) in company activities through the variable of mobile collaboration technologies, (2) Engineering asset management with a specific focus on the most critical process – asset maintenance, and (3) Comprehensive framework that meet all requirements (technological, organisational and people perspectives).

References

1. CIEAM, Centre for Integrated Engineering Asset Management, http://www.cieam.com/site/34/publications
2. Tsang, A.H.C.: Strategic Dimension of Maintenance Management. Journal of Quality in Maintenance Engineering 8(1), 7–39 (2002)
3. Hodkiewicz, M.R., Pascual, R.: Education in Engineering Asset Management Current Trends and Challenges. Presented at The International Physical Asset Management Conference (2006)
4. Snitkin, S.: Collaborative Asset Lifecycle Management Vision and Strategies, Research Report, ARC Advisory Group, Boston, USA (2003)
5. Camacho, J., Galicia, L., Gonzales, V.M., Favela, J.: MobileSJ: Managing Multiple Activities in Mobile Collaborative Working Environments. International Journal of e-Collaboration 4(1), 61–73 (2008)
6. Laszkiewics, M.: Collaborative Maintenance: a Strategy to Help Manufactures Become Learn, Mean, and Agile. Plant Engineering 57(9), 30–36 (2003)
7. Ferrario, M.A., Smyth, B.: Distributing case-base maintenance: the collaborative maintenance approach. Computational Intelligence 17(2), 315–330 (2001)
8. Besten, M., Dalle, J.M., Galia, F.: Collaborative Maintenance in Large Open-Source Projects. International Federation for Information Processing 23, 233–244 (2006)
9. Rein, G.L.: Collaboration Technology for Organizational Design. In: Proceedings of the 26th Hawaii International Conference (1993)
10. Márquez, A.C.: The Maintenance Management Framework. Springer, London (2007)
11. Sokianos, N., Druke, H., Toutatoui, C.: Lexikon Produktions Management, Landsberg, Germany (1998)
12. Tam, S.B., Price, J.W.H.: Optimisation Framework for Asset Maintenance investment. Monash Business Review 2(3), 1–10 (2006)
13. Zhang, Z., Li, Z., Huo, Z.: CMMS and its Application in Power System. International Journal of Power & Energy Systems 26(1), 75–82 (2006)

14. Olszwesky, R.: RCM success starts with CMMS, http://www.maintenanceworld.com/Articles/reliabilityweb/RCM_Success_CMMS.pdf
15. Liang, T.P., Huang, C.W., Yeh, Y.H.: Adoption of Mobile Technology in Business: a Fit-Viability Model. Industrial Management & Data Systems 107, 1154–1169 (2007)
16. Monostori, L., Vancza, J., Kumara, S.R.T.: Agent-Based Systems for Manufacturing. Annals of the CIRP 55, 697–720 (2006)
17. Linstone, H.A.: Decision Making for Technology Executives: Using Multiple Perspectives to Improve Performance. Artech House Publisher, London (1999)
18. Klein, H.K., Myers, M.D.: A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems. MIS Quarterly 23(1), 67–93 (1999)
19. Deetz, S.: Describing Differences in Approaches to Organization Science: Re-thinking Burrell and Morgan and their legacy. Organization Science 7(2), 191–207 (1996)
20. Orlikowski, W.J., Baroudi, J.J.: Studying Information Technology in Organizations: research approaches and assumptions. Information Systems Research 2(1), 1–28 (1991)
21. Greene, J.C., Caracelli, V.J.: Making Paradigmatic Sense of Mixed Methods Practice. In: Tashakkori, A., Teddlie, C. (eds.) Handbook of Mixed Methods in Social and Behavioral Research. Sage Publications, Thousand Oaks (2003)
22. Dalkey, N.C., Helmer, O.: An Experimental Application of the Delphi Method to the Use of Experts. Management Science 9, 458–467 (1963)
23. Linstone, H.A., Turoff, M.: The DelphiMethod: Techniques and Applications. Addison-Wesley, London (1975)

A Genetic Algorithm for Power-Aware Virtual Machine Allocation in Private Cloud

Nguyen Quang-Hung, Pham Dac Nien, Nguyen Hoai Nam,
Nguyen Huynh Tuong, and Nam Thoai

Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology
268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam
{hungnq2,htnguyen,nam}@cse.hcmut.edu.vn,
{50801500,50801308}@stu.hcmut.edu.vn

Abstract. Energy efficiency has become an important measurement of scheduling algorithm for private cloud. The challenge is trade-off between minimizing of energy consumption and satisfying Quality of Service (QoS) (e.g. performance or resource availability on time for reservation request). We consider resource needs in context of a private cloud system to provide resources for applications in teaching and researching. In which users request computing resources for laboratory classes at start times and non-interrupted duration in some hours in prior. Many previous works are based on migrating techniques to move online virtual machines (VMs) from low utilization hosts and turn these hosts off to reduce energy consumption. However, the techniques for migration of VMs could not use in our case. In this paper, a genetic algorithm for power-aware in scheduling of resource allocation (GAPA) has been proposed to solve the static virtual machine allocation problem (SVMAP). Due to limited resources (i.e. memory) for executing simulation, we created a workload that contains a sample of one-day timetable of lab hours in our university. We evaluate the GAPA and a baseline scheduling algorithm (BFD), which sorts list of virtual machines in start time (i.e. earliest start time first) and using best-fit decreasing (i.e. least increased power consumption) algorithm, for solving the same SVMAP. As a result, the GAPA algorithm obtains total energy consumption is lower than the baseline algorithm on simulated experimentation.

1 Introduction

Cloud computing [7], which is popular with pay-as-you-go utility model, is economy driven. Saving operating costs in terms of energy consumption (Watts-Hour) for a cloud system is highly motivated for any cloud providers. Energy-efficient resource management in large-scale datacenter is still challenge [1][13][9][5]. The challenge of energy-efficient scheduling algorithm is trade-off between minimizing of energy consumption and satisfying demand resource needs on time and non-preemptive. Resource requirements depend on the applications and we are

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 183–191, 2013.

interested in virtual computing lab, which is a cloud system to provide resources for teaching and researching.

There are many studies on energy efficient in datacenters. Some studies proposed energy efficient algorithm that are based on processor speed scaling (assumption that CPU technology supports dynamic scaling frequency and voltage (DVFS)) [1][13]. Some other studies proposed energy efficient by scheduling for VMs in virtualized datacenter [9][5]. A. Beloglazov et al. [5] presents the Modified Best-Fit Decreasing (MBFD) algorithm, which is best-fit decreasing heuristic, for power-aware VM allocation and adaptive threshold-based migration algorithms to dynamic consolidation of VM resource partitions. Goiri, . et al. [9] presents score-based scheduling, which is hill-climbing algorithm, to place each VM onto which physical machine has the maximum score. However, the challenge is still remain. These previous works did not concern on satisfying demand resource needs on time (i.e. VM starts at a specified start time) and non-preemptive, in addition to both MBFD and score-based algorithms do not find an optimal solution for VM allocation problem.

In this paper, we introduce our static virtual machine allocation problem (SVMAP). To solve the SVMAP, we propose the GAPA, which is a genetic algorithm to find an optimal solution for VM allocation. On simulated experimentation, the GAPA discovers a better VM allocation (means lower energy consumption) than the baseline scheduling algorithm for solving same SVMAP.

2 Problem Formulation

2.1 Terminology, Notation

We describe notation that is used in this paper as following:

- VM_i: the i-th virtual machine
- M_j: the j-th physical machine
- ts_i: start time of the VM_i
- pe_i: number of processing elements (e.g. cores) of the VM_i
- PE_j: number of processing elements (e.g. cores) of the M_j
- $mips_i$: total required MIPS (Millions Instruction Per Seconds) of the VM_i
- $MIPS_j$: total capacity MIPS (Millions Instruction Per Seconds) of the M_j
- d_i: duration time of the VM_i, units in seconds
- $P_j(t)$: power consumption (Watts) of a physical machine M_j
- $r_j(t)$: set of indexes of virtual machines that is allocated on the M_j at time t

2.2 Power Consumption Model

In this section, we introduce factors to model the power consumption of single physical machine. Power consumption (Watts) of a physical machine is sum of total power of all components in the machine. In [8], they estimated power consumption of a typical server (with 2x CPU, 4x memory, 1x hard disk drive,

2x PCI slots, 1x mainboard, 1x fan) in peak power (Watts) spends on main components such as CPU (38%), memory (17%), hard disk drive (6%), PCI slots (23%), mainboard (12%), fan (5%). Some papers [8] [4] [6] [5] prove that there exists a power model between power and resource utilization (e.g. CPU utilization). We assume that power consumption of a physical machine ($P(.)$) is linear relationship between power and resource utilization (e.g. CPU utilization) as [8][4][6][5]. The total power consumption of a single physical server ($P(.)$) is:

$$P(U_{cpu}) = P_{idle} + (P_{max} - P_{idle})U_{cpu}$$

$$U_{cpu}(t) = \sum_{c=1}^{PE_j} \sum_{i \in r_j(t)} \frac{mips_{i,c}}{MIPS_{j,c}}$$

In which:

- $U_{cpu}(t)$: CPU utilization of the physical machine at time t, $0 \leq U_{cpu}(t) \leq 1$
- P_{idle}: the power consumption (Watt) of the physical machine in idle, e.g. 0% CPU utilization
- P_{max}: the maximum power consumption (Watt) of the physical machine in full load, e.g. 100% CPU utilization
- $mips_{i,c}$: requested MIPS of the c-th processing element (PE) of the VM_i
- $MIPS_{j,c}$: Total MIPS of the c-th processing element (PE) on the physical machine M_j

The number of MIPS that a virtual machine requests can be changed by its running application. Therefore, the utilization of the machine may also change over time due to application. We link the utilization with the time t. We re-write the total power consumption of a single physical server ($P(.)$) with $U_{cpu}(t)$ as:

$$P(U_{cpu}(t)) = P_{idle} + (P_{max} - P_{idle})U_{cpu}(t)$$

and total energy consumption of the physical machine (E) in period time $[t_0, t_1]$ is defined by:

$$E = \int_{t_0}^{t_1} P(U_{cpu}(t))dt$$

2.3 Static Virtual Machine Allocation Problem (SVMAP)

Given a set of n virtual machines $\{VM_i(pe_i, mips_i, ts_i, d_i)|i = 1, ..., n\}$ to be placed on a set of m physical parallel machines $\{M_j(PE_j, MIPS_j)|j = 1, ..., m\}$. Each virtual machine VM_i requires pe_i processing elements and total of $mips_i$ MIPS, and the VM_i will be started at time (ts_i) and finished at time ($ts_i + d_i$) without neither preemption nor migration in its duration (d_i). We do not limit resource type on CPU. We can extend for other resource types such as memory, disk space, network bandwidth, etc.

Algorithm 1. GAPA Algorithm

Start: Create an initial population randomly for s chromosomes (with s is population size)
Fitness: Calculate evaluation value of each chromosome respectively in given population.
New population: Create a new population by carrying out follows the steps:
Selection: Choose the two individual parents from current population based on value of evaluation.
Crossover: By using crossover probability, we create new children via modifying chromosome of parents.
Mutation: With mutation probability, we will mutate at some position on chromosome.
Accepting: Currently, new children will be a part of the next generation.
Replace: Go to the next generation by assigning the current generation to the next generation.
Test: If stop condition is satisfied then this algorithm is stopped and returns individual has the highest evaluation value. Otherwise, go to next step.
Loop: Go back the Fitness step.

We assume that every physical machine M_j can host any virtual machine, and its power consumption model ($P_j(t)$) is proportional to resource utilization at a time t, e.g. power consumption has a linear relationship with resource utilization (e.g. CPU utilization) [8][2][5].

The objective scheduling is minimizing energy consumption in fulfillment of maximum requirements of n VMs.

2.4 The GAPA Algorithm

The GAPA, which is a kind of Genetic Algorithm (GA), solves the SVMAP. The GAPA performs steps as in the Algorithm 1.

In the GAPA, we use a tree structure to encode chromosome of an individual. This structure has three levels:

Level 1: Consist of a root node that does not have significant meaning.
Level 2: Consist of a collection of nodes that represent set of physical machines.
Level 3: Consist of a collection of nodes that represent set of virtual machines.

With above representation, each instance of tree structure will show that an allocation of a collection of virtual machines onto a collection of physical machines. The fitness function will calculate evaluation value of each chromosome as in the Algorithm 2.

Algorithm 2. Construct fitness function

```
powerOfDatacenter := 0
For each host ∈ collection of hosts do
    utilizationMips := host.getUtilizationOfCpu()
    powerOfHost := getPower (host, utilizationMips)
    powerOfDatacenter := powerOf Datacenter + powerOfHost
End For
Evaluation value (chromosome) := 1.0 / powerOfDatacenter
```

3 Experimental Study

3.1 Scenarios

We consider on resource allocation for virtual machines (VMs) in private cloud that belongs to a college or university. In a university, a private cloud is built to provide computing resource for needs in teaching and researching. In the cloud, we deploy installing software and operating system (e.g. Windows, Linux, etc.) for practicing lab hours in virtual machine images (i.e. disk images) and the virtual machine images are stored in some file servers. A user can start, stop and access VM to run their tasks. We consider three needs as following:

i A student can start a VM to do his homework.
ii A lecturer can request a schedule to start a group of identical VMs for his/her students on lab hours at specified start time and in prior. The lab hours requires that the group of VMs will start on time and continue in spanning some time slots (e.g. 90 minutes).
iii A researcher can start a group of identical VMs to run his/her parallel application.

3.2 Workload and Simulated Cluster

We use workload from one-day of our university's schedule for laboratory hours on six classes in the Table 1. The workload is simulated by total of 211 VMs and 100 physical machines (hosts).

We consider there are two kind of servers in our simulated virtualized datacenter, which includes two power consumption models of two power model of the IBM server x3250 (1 x [Xeon X3470 2933 MHz, 4 cores], 8GB) and another power model of the Dell Inc. PowerEdge R620 (1 x [Intel Xeon E5-2660 2.2 GHz, 16 cores], 24 GB) server with 16 cores in the Table 2. The baseline scheduling algorithm (BFD), which sorts list of virtual machines in start time (i.e. earliest start time first) and using best-fit decreasing (i.e. least increased power consumption, for example MBFD [5]), will use four IBM servers to allocate for 16 VMs (each VM requests single processing element). Our GAPA can finds a better VM allocation (lesser energy consumption) than the minimum increase of power consumption (best-fit decrease) heuristic in our experiments. In this example, our GAPA will choose one Dell server to allocate these 16 VMs. As a result, our GAPA consumes less total energy than the BFD does.

Table 1. Workload of a university's one-day schedule

Day	Subject	Class ID	Group ID	Students	Lab. Time	Duration (sec.)
6	506007	CT10QUEE	QT01	5	—456———-	8100
6	501129	CT11QUEE	QT01	5	123———-	8100
6	501133	DUTHINH6	DT04	35	123———-	8100
6	501133	DUTHINH5	DT01	45	—456———-	8100
6	501133	DUTHINH5	DT02	45	—456———-	8100
6	501133	DUTHINH6	DT05	35	123———-	8100
6	501133	DUTHINH6	DT06	41	123———-	8100

3.3 Experiments

We show results from the experiments in the Table 3 and Figure 1. We use a popular simulated software for a virtualized datacenter is the CloudSim [14][6] to simulate our virtualized datacenter and the workload. The GAPA is a VM allocation algorithm that is developed and integrated into the CloudSim version 3.0.

On simulated experimentation, we have total energy consumptions of both the BFD and the GAPA algorithms are 16.858KWh and average of 13.007KWh respectively. We conclude that the energy consumption of the BFD algorithm is higher than the energy consumption of GAPA algorithm is approximately 130%. In case of the GAPA, these GAPA use the probability mutation is 0.01 and size of population is 10, number of generations is {500, 1000}, probability of crossover is {0.25, 0.5, 0.75}.

Table 2. Two power models of (i) the IBM server x3250 (1 x [Xeon X3470 2933 MHz, 4 cores], 8GB) [16] and (ii) the Dell Inc. PowerEdge R620 (1 x [Intel Xeon E5-2660 2.2 GHz, 16 cores], 24 GB) [15]

Utilization	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
IBM x3250	41.6	46.7	52.3	57.9	65.4	73.0	80.7	89.5	99.6	105.0	113.0
Dell R620	56.1	79.3	89.6	102.0	121.0	132.0	149.0	171.0	195.0	225.0	263.0

4 Related Works

B. Sotomayor et al. [12] proposed a lease-based model and First-Come-First-Serve (FCFS) and backfilling algorithms to schedule best-effort, immediate and advanced reservation jobs. The FCFS and backfilling algorithms consider only performance metric (e.g. waiting time, slowdown). To maximize performance, these scheduling algorithms tend to choose free load servers (i.e. highest-ranking scores) when allocates a new lease. Therefore, a lease with single VM can be allocated on big, multi-core physical machine. This way could be waste energy, both of the FCFS and backfilling does not consider on the energy efficiency.

Table 3. Total energy consumption (KWh) of running: (i) earliest start time first with best-fit decreasing (BFD); (ii) GAPA algorithms. These GAPA use the probability mutation of 0.01 and size of population of 10. N/A means not available.

Algorithms	VMs	Hosts	GA's Generations	GA's Prob. of Crossover	Energy (KWh)	BFD/GAPA
BFD	211	100	N/A	N/A	16.858	1
GAPA_P10_G500_C25	211	100	500	0.25	13.007	1.296
GAPA_P10_G500_C50	211	100	500	0.50	13.007	1.296
GAPA_P10_G500_C75	211	100	500	0.75	13.007	1.296
GAPA_P10_G1000_C25	211	100	1000	0.25	13.007	1.296
GAPA_P10_G1000_C50	211	100	1000	0.50	13.007	1.296
GAPA_P10_G1000_C75	211	100	1000	0.75	13.007	1.296

S. Albers et al. [1] reviewed some energy efficient algorithms which are used to minimize flow time by changing processor speed adapt to job size. G. Laszewski et al. [13] proposed scheduling heuristics and to present application experience for reducing power consumption of parallel tasks in a cluster with the Dynamic Voltage Frequency Scaling (DVFS) technique. We did not use the DVFS technique to reduce energy consumption on datacenter.

Some studies [9][3][5] proposed algorithms to solve the virtual machine allocation in private cloud to minimize energy consumption. A. Beloglazov et al. [3][5] presented a best-fit decreasing heuristic on VM allocation, named MBFD, and VM migration policies under adaptive thresholds. The MBFD tends to allocate a VM to such as active physical machine that would take the minimum increase of power consumption (i.e. the MBFD prefers a physical machine with minimum power increasing). However, the MBFD cannot find an optimal allocation for all VMs. In our simulation, for example, the GAPA can find a better VM allocation (lesser energy consumption) than the minimum increase of power consumption (best-fit decrease) heuristic in our experiments. In this example, our GAPA will choose one Dell server to allocate these 16 VMs. As a result, our GAPA consumes less total energy than the best-fit heuristic does.

Another study on allocation of VMs [9] developed a score-based allocation method to calculate scores matrix of allocations of m VMs to n physical machines. A score is sum of many factors such as power consumption, hardware and software fulfillment, resource requirement. These studies are only suitable for service allocation, in which each VM will execute a long running, persistent application. We consider each user job has a limited duration time. In addition to, our GAPA can find an optimal schedule for the static VM allocation problem on single objective is minimum energy consumption.

In a recently work, J. Kolodziej et al. [10] presents evolutionary algorithms for energy management. None of these solutions solves same our SVMAP problem.

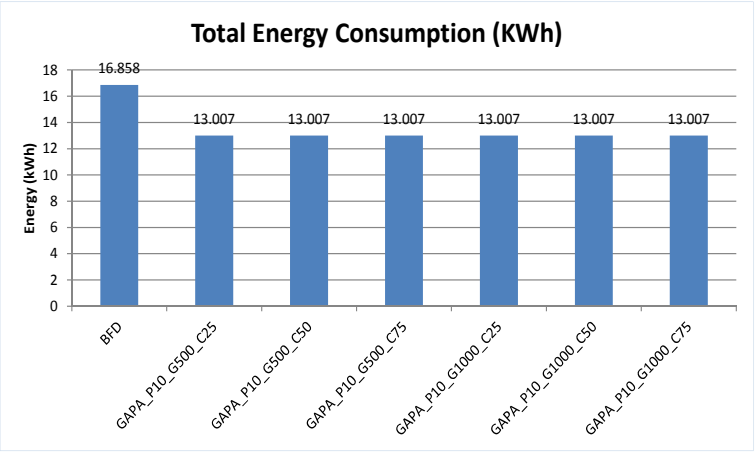


Fig. 1. The total energy consumption (KWh) for earliest start time first with best-fit decrease (BFD), GAPA algorithms

5 Conclusions and Future Works

In a conclusion, a genetic algorithm can apply to the static virtual machine allocation problem (SVMAP) and brings benefit in minimize total energy consumption of computing servers. On simulation with workload of one-day lab hours in university, the energy consumption of the baseline scheduling algorithm (BFD) algorithm is higher than the energy consumption of GAPA algorithm is approximately 130%. Disadvantage of the GAPA algorithm is longer computational time than the baseline scheduling algorithm.

In the future work, we concern methodology to reduce computational time of the GAPA. We also concern some other constraints, e.g. deadline of jobs. We also study on migration policies and history-based allocation algorithms.

References

1. Albers, S., Fujiwara, H.: Energy-efficient algorithms. ACM Review 53(5), 86–96 (2010), doi:10.1145/1735223.1735245
2. Barroso, L.A., Hölzle, U.: The Case for Energy-Proportional Computing, vol. 40, pp. 33–37. ACM (2007), doi:10.1109/MC.2007.443
3. Beloglazov, A., Buyya, R.: Energy Efficient Resource Management in Virtualized Cloud Data Centers. In: Proceedings of the 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, pp. 826–831 (2010), doi:10.1109/CCGRID.2010.46
4. Beloglazov, A., Buyya, R.: Adaptive Threshold-Based Approach for Energy-Efficient Consolidation of VMs in Cloud Data Centers. ACM (2010)
5. Beloglazova, A., Abawajy, J., Buyya, R.: Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing. FGCS 28(5), 755–768 (2012), doi:10.1016/j.future.2011.04.017

6. Beloglazov, A., Buyya, R.: Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers. In: Concurrency and Computation: Practice and Experience, Concurrency Computat.: Pract. Exper., pp. 1–24 (2011), doi: 10.1002/cpe
7. Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J., Brandic, I.: Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. FGCS 25(6), 599–616 (2009), doi:10.1016/j.future.2008.12.001
8. Fan, X., Weber, W.-D., Barroso, L.A.: Power provisioning for a warehouse-sized computer. In: Proceedings of the 34th Annual International Symposium on Computer Architecture, pp. 13–23. ACM (2007), doi:10.1145/1273440.1250665
9. Goiri, J.F., Nou, R., Berral, J., Guitart, J., Torres, J.: Energy-aware Scheduling in Virtualized Datacenters. In: IEEE International Conference on Cluster Computing, CLUSTER 2010, pp. 58–67 (2010)
10. Kołodziej, J., Khan, S.U., Zomaya, A.Y.: A Taxonomy of Evolutionary Inspired Solutions for Energy Management in Green Computing: Problems and Resolution Methods. In: Kolodziej, J., Khan, S.U., Burczynski, T., et al. (eds.) Advances in Intelligent Modelling and Simulation. SCI, vol. 422, pp. 215–233. Springer, Heidelberg (2012)
11. Sotomayor, B., Keahey, K., Foster, I.: Combining batch execution and leasing using virtual machines. In: Proceedings of the 17th International Symposium on High Performance Distributed Computing - HPDC 2008, pp. 87–96. ACM (2008), doi: 10.1145/1383422.1383434
12. Sotomayor, B.: Provisioning Computational Resources Using Virtual Machines and Leases, PhD Thesis submited to The University of Chicago, US (2010)
13. Laszewski, G.V., Wang, L., Younge, A.J., He, X.: Power-aware scheduling of virtual machines in DVFS-enabled clusters. In: 2009 IEEE International Conference on Cluster Computing and Workshops, pp. 368–377 (2009), doi:10.1109/CLUSTR.2009.5289182
14. Calheiros, R.N., Ranjan, R., Beloglazov, A., De Rose, C.A.F., Buyya, R.: CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Software: Practice and Experience 41(1), 23–50 (2011)
15. SPECpower ssj2008 results for Dell Inc. PowerEdge R620 (Intel Xeon E5-2660, 2.2 GHz)
`http://www.spec.org/power_ssj2008/results/res2012q2/power_ssj2008-20120417-00451.html`
(last accessed November 29, 2012)
16. SPECpower ssj2008 results for IBM x3250 (1 x [Xeon X3470 2933 MHz, 4 cores], 8GB).
`http://www.spec.org/power_ssj2008/results/res2009q4/power_ssj2008-20091104-00213.html`
(last accessed November 29, 2012)

Cloud-Based E-Learning: A Proposed Model and Benefits by Using E-Learning Based on Cloud Computing for Educational Institution

Nungki Selviandro and Zainal Arifin Hasibuan

Digital Library and Distance Learning Laboratory
Faculty of Computer Science Universitas Indonesia
nungki.selviandro@ui.ac.id, zhasibua@cs.ui.ac.id

Abstract. The increasing research in the areas of information technology have a positive impact in the world of education. The implementation of e-learning is one of contribution from information technology to the world of education. The implementation of e-learning has been implemented by several educational institutions in Indonesia. E-Learning provides many benefits such as flexibility, diversity, measurement, and so on. The current e-learning applications required large investments in infrastructure systems regardless of commercial or open source e-learning application. If the institution tended to use open source e-learning application it would need more cost to hire professional staff to maintain and upgrade the e-learning application. It can be challenging to implement e-learning in educational institutions. Another problem that can arise in the use of e-laerning trend today is more likely to institution building their own e-learning system itself. If two or more institutions are willing to build and use an e-learning so they can minimize the expenditure to develop the system and share learning materials more likely happened. This paper discuss the current state and challenges in e-learning and then explained the basic concept and previous proposed architectures of cloud computing. In this paper authors also proposed a model of cloud-based e-learning that consists of five layer, namely: (1) infrastructure layer; (2) platform layer; (3) application layer; (4) access layer; and (5) user layer. In addition to this paper we also illustrated the shift paradigm from conventional e-learning to cloud-based e-learning and described the expected benefits by using cloud-based e-learning.

Keywords: E-Learning, Cloud Computing, Cloud-Based E-Learning.

1 Introduction

Nowadays e-learning widely use by educational institutions for supporting their learning process and provide anytime service for learners to access learning material and information. The implementation of e-learning has been implemented by several educational institutions in Indonesia. E-Learning provides many benefits such as flexibility, diversity, measurement, and others [11], even though its implementation still exist many difficulties. The main problem experienced when to start applying e-learning is

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 192–201, 2013.

the high initial cost or in other words is the economic factor [2]. It is becoming a major focus for the institutions that will implementing e-learning. Institutions are categorized as low budget certainly be very difficult to implement e-learning, even if the institution has an adequate budget also expects a minimal budget that can be spent to implement e-learning. The inadequate infrastructure becomes a major problem in the implementation of e-learning. Institutions that want to implement e-learning difficulties in the procurement of server/PC, storage, and network [3]. Besides provide infrastructure The next issue is human resources, not all of the institutions have the professional staff for designing, developing systems to manage e-learning, in addition to the growing applying of e-learning as well is required expert in designing teaching materials commonly known as the instructional designer. It is also a consideration in the implementation of e-learning for each institution to implement it. Because institutions will also estimate the cost to employ them in order to provide specifically for e-learning systems.

Along with the development of the IT world, cloud computing is gradually become the new paradigm of innovation in the IT world, cloud computing is a computing services that can be used through the Internet in accordance with the needs of users with little interaction between service providers and users. Cloud computing technology as well described as a computing resource that provides a highly scalable as external services through the Internet. Therefore, cloud computing can be considered as an alternative to minimize the cost of infrastructure and human resources for development and maintenance process of e-learning systems [4].

In this paper the author will discuss the current state and challenges of e-learning as well as basic concepts of cloud cover and the implementation of the service model, and the author will discuss some of the architecture of cloud-based e-learning that has been proposed by previous researchers. In addition the author will introduce the model proposed in the implementation of e-learning in the cloud environment as an alternative to conventional e-learning implementations are widely used in educational institutions today. The author will also explain the expected benefit by adopting the model of cloud-based e-learning.

2 E-Learning: Current State and Challenges

E-Learning is an internet-based learning process which aims to support conventional learning process that using internet technology and will not replace traditional education method [11]. Usually, e-learning systems are based on client-server architecure and web-based technology [17]. This architecture has some limitation so that e-learning can not be used to its full potential, because has some limitation such as lack of interoperability and accessibility. Based on previous study, in order to solve interoperability issue, the use of web services has been implemented by several previous researchers as practiced by Grewal et al. (2005), Pankratius et al. (2004), Xu et al. (2003). The using of web service has successfully answered the issue of interoperability from e-learning with focusing on selecting and combining the learning objects [5].

With the development of mobile technology make e-learning is increasingly being used. The use of mobile technology in the implementation of e-learning is commonly known as mobile learning. There are many definitions of mobile learning, one of which is mobile learning defined by Lan & Sie (2010) as a learning model that enables participants to achieve the teaching learning materials anywhere and anytime using mobile technology and the internet. This definition may mean that mobile learning could include mobile phones, smartphones, personal digital assistants (PDAs) and their peripherals. Using mobile learning can help address the issue of accessibility in accessing the e-learning system.

Nowadays the use of e-learning applications can be based on commercial products or from open source. The advantage of using commercial products are the implementation time is quick due to technical support from the vendor and there will be ongoing maintenance cost. The disadvantage of using commercial e-learning applications are the initial cost of procuring commercial e-learning software is very high and there will the cost of insfrastructure [4]. Open source e-learning applications widely used in university. The initial cost of e-learning software is very low, but there still need expensive investment for the infrastructure and need more cost to hire professional staff for maintaining and upgrading the e-learning applications.

Based on the above phenomenon, the current e-learning applications required large investments in infrastructure systems regardless of commercial or open source. If the institution tended to use open source e-learning application it would need more cost to hire professional staff to maintain and enhance the e-learning application. It can be challenging to implement e-learning in educational institutions. Another problem that can arise in the use of e-laerning trend today is more likely to institution building their own e-learning system itself. If two or more institutions are willing to build and use an e-learning so they can minimize the expenditure to develop the system and share learning materials more likely happened.

3 Cloud Computing

Cloud Computing is a new paradigm to organize and manage ICT resources. There are various definitions of cloud computing, one of which is the definition according to The National Institute of Standards and Technology (NIST) which defines cloud computing as "*model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction*" [22]. Generally speaking, the cloud computing service model consists of three layers [5], among others: (1) Software as a Service (SaaS); (2) Platform as a service (PaaS); (3) Infrastructure as a service (IaaS) [6].

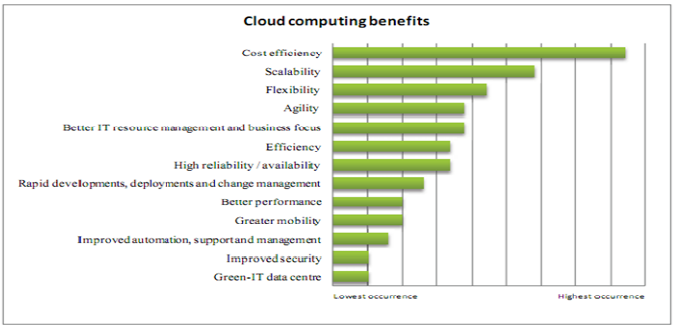


Fig. 1. The Advantages of implementing *Cloud Computing* [8]

In practice, cloud computing has four implementation models where each model has certain characteristics [7], among others: (1) Private, the model is aimed at an organization where cloud operations are managed by a third party or the organization itself; (2) Public, service on this model is intended for the general public or the industry in which the various services provided by the cloud computing service provider organization (3) Community, this model is managed by several organizations that form a community of practice in which the operations are managed by the community with the division of tasks particular; (4) Hybrid, this model is a combination of various models existing cloud distribution. Typically, this is done with a combination of specific purposes where there is an attachment for example: technological standards and data ownership.

The implementation of cloud computing is growing due to have advantages such as that illustrated in Figure 1, it can be seen that the efficiency cost is a major indicator of the advantages possessed by Cloud Computing. Cost efficiency can be realized due to several components such as the cost of financing the purchase of infrastructure and application development and operational expenses (management fees and maintenance) can be reduced. Cloud Computing on the implementation can be categorized two types, namely free service or pay per use (pay-as-you-go services) [12], users will only be charged when using services from providers the service.

4 Cloud-Based E-Learning Architecture

In this chapter will discussed about the previous cloud-based architecture that developed by former researcher in this area and also the proposed cloud-based e-learning archictecture from this paper. In addition for this chapter also will be discussed the contribution from the proposed architecture.

4.1 Previous Cloud-Based E-Learning Architecture

The research of e-learning in the cloud environment have been carried out by previous researchers, such as those conducted by Chuang, Chang, and Sung (2011), Dong et al (2009), Vishwakarma & Narayanan (2011), Pocatilu (2010) and Ghazizadeh (2012). Research on the application of e-learning in a cloud environment is one form of cloud services education services. There are several architectural cloud-based e-learning have been proposed by previous researcher. In this paper will discuss three architectural cloud-based e-learning, such as architecture proposed by Phankokkruad (2012), Wang, Pai, & Yen (2011), and Masud & Huang (2012).

Phankokkruad (2012) proposed e-learning architecture based on cloud computing consists of three layers: (1) infrastructure layer, (2) platform (middle) layer, and, (3) application layer. Infrastructure layer is a hardware layer that supplies the computing and storage capacity for the higher level and this layer, which is used as e-learning and software virtualization tehcnologies, ensures the stability and reliability of the infrastructure. The second layer is Platform layer, this layer is a middle layer consisting middleware that is Web service. It is used for providing the learning resources as a service. This layer consists of two modules: item classification module (ICM) and course selection module (CSM). They are used for accessing the items from the item bank and selecting suitable learning content from the content database. The third layer is Application layer which is responsible for interface provision for the students.

Not much difference can be inferred from the comparison of the architecture delivered by Phankokkruad (2012) and Wang, Pai, & Yen (2011). They proposed an architecture of e-learning-based cloud computing consists of three layers, namely: (1) infrastructure layer, (2) middleware layer, and, (3) application layer.

The first layer is infrastructure layer. It is employed as the e-learning resource pool that consists of hardware and software virtualization technologies to ensure the stablility and reliability of the infrastructure. This layer also supplies the computing and storage capacity for the higher level. The second layer is middleware layer. It focuses in providing a sharable platform consisting of two modules: CNRI's (Corporation for National Research Initiatives) Handler System Module and Metadata Transformation System Module. The final layer is application layer. At this layer, cloud computing provides convenient access to the e-learning resources.

The next architecture proposed by Masud & Huang (2012) consists of five layers. The First layer is infrastructure layer. It is composed of information infrastructure and teaching resources. Information infrastructure contains internet/intranet, system software, information management system and some common hardware. Teaching resources stored up mainly in traditional teaching model and distributed in different departments and domain. The second layer is software resource layer. This layer is composed by operating system and middleware. A variety of software resources are integrated through middleware technology to provide a unified interface for software developers to develop applications and embed them in the cloud. The third layer is resource management layer. In order to effectuate on demand free flow and distribution of software over various hardware resources, this layer utilizes integration of virtualization and cloud computing scheduling strategy. The fourth layer is service

layer. This layer has three levels of services namely, SaaS, PaaS, and IaaS. In SaaS, cloud computing service is provided to customers, contrasting to traditional software, cloud customers use software via the internet without any need to purchase, maintain, and upgrade. They simply to pay a monthly fee. The last layer is application layer. This layer is a specific layer consisting of applications of integrated teaching resources, including interactive courses and the teaching resources sharing. The teaching resources include teaching material, teaching information, as well as the full sharing human resources.

4.2 Proposed Cloud-Based E-Learning Architecture

In this paper we propose the architecture that we have designed by modifying previous architectures that we used as references. Our proposed architecture consists of five layers (as shown in Figure 5), namely: (1) infrastructure layer; (2) platform layer; (3) application layer; (4) access layer; and (5) user layer.

First layer is infrastructure layer. This layer contains architecture supporting infrastructure, such as: Cloud platform, virtual machine, virtual repositories and physical infrastructure such as servers, network devices, storage, buildings and other physical facilities. The infrastrucuter layer shares IT infrstructure resources and connects the system huge system pool together to provide services. Cloud computing enable the hardware layer to run more like the internet, to make the hardware resources shared and accessed the data resources in secure and scalable way. The second layer is platform layer. In this layer running the operating system where e-learning application will be running. Besides the operating system, this layer also consists of variety of software that support the application layer so that it can run properly. The third layer

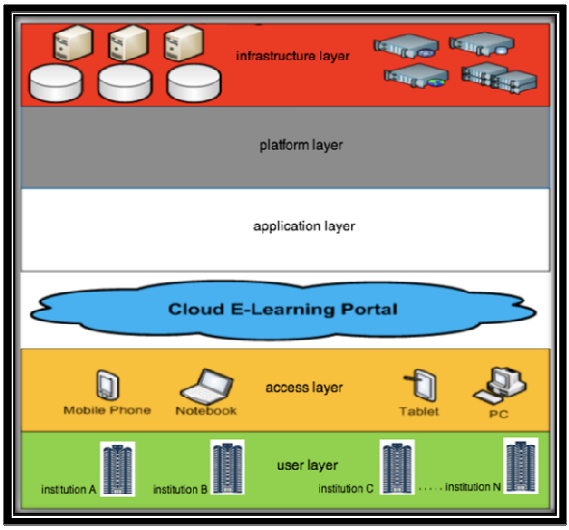


Fig. 2. Proposed Cloud E-Learning Architecture

is application layer. This layer is a specific e-learning application that is utilized for sharing learning resources and interaction among users that includes synchronous or asynchronous discussion and chatting. We added the access layer in our architecture. This access layer is the fourth layer in our proposed architecture. This layer is in charge of managing access to cloud e-learning services which is available on the architecture such as: types of access devices and presentation models. This study adopts the concept of multi-channel access which enables a variety of available services that accessible through a variety of devices (such as mobile phones, smartphones, computer, etc.) and a variety of presentation models (such as mobile applications, desktop applications, and others) [21]. The purpose of the adoption of this concept is to increase the availability of devices that access the cloud service e-learning can be found in the architecture used untrammeled access devices. Besides the addition of the access layer, the architecture we propose the user layer consists of various educational institutions.

5 Conventional E-Learning towards Cloud-Based E-Learning

Based on [8] which is also illustrated in Figure 1, the main advantages of the adoption of cloud computing is efficient in terms of cost, this is an interesting point of view that can be adapted to develop e-learning based on cloud computing. Conventional e-learning commonly used by the university developed by the university itself (shown in Figure 6) tend to cause lots of problems such as time to designing e-learning systems will be developed, costs for infrastructure, selecting commercial or open source e-learning platform , the cost to hire professional staff to maintain and upgrade the system of e-learning, and so on. This process is more likely need more time.

By introducing cloud computing adopted by e-learning, as shown in Figure 6, institutions can use a single e-learning based on the cloud provided by a cloud provider of e-learning. This model can reduce the initial costs incurred by the institution for the implementation of e-learning by using cloud computing services, because institutions do not need to pay for the purchase of infrastructure, both in terms of procurement of servers and storage. With cloud computing, as an institution of the client can rent the infrastructure to cloud computing service providers [7]. Likewise with the human resources for the development stage, the cloud environment of e-learning has been provided by the cloud service provider, as well as maintenance of the e-learning [9].

The paradigm shift in the implementation of e-learning is an innovation that can help any institution in implementing e-learning. In general, the implementation conventional e-learning, e-learning web-based design, system development and maintenance as well as by internal governance institutions [10]. It had a lot of problems, both in terms of flexibility, scalability, and accessibility [5] [7] [11]. According to [12] are discussed in [3], one of the main important features that can be presented in the use of e-learning in the cloud is scalability, which allows virtualization provide infrastructure layer provided by the cloud service provider. Virtualization helps solve the problem of the physical barriers that are generally inherent in the lack of resources and infrastructure to automate the management of these resources as if they were a single entity through hypervisor technologies such as virtual machine (VM).

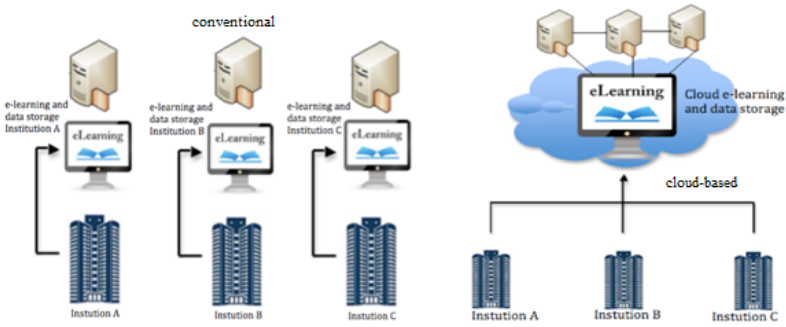


Fig. 3. Conventional E-Learning Towards Cloud-Based E-Learning

6 Expected Benefits

The expected advantages by adopting the cloud-based e-learning model are as follows: (1) Large capacity, this criteria could adress on-demand self service characteristic from could computing. Large scale storage in cloud environments provide advantages to the consumer to determine the storage capacity they intend to use that are adjusted to their needs and capabilities of the institution as a consumer of cloud-based e-learning; (2) Short implementation process, by using cloud-based e-learning services, educational instition could minimize their expenditure to develop the e-learning system and shorten the implementation process because the e-learning system already developed and maintained by the cloud e-learning provider; (3) High Availability, by utilizing large storage and high performance computing power, cloud e-learning could provide a high quality of service. This may happened because of the support system that supports cloud e-learning can detect the node failure and can be immediately diverted to another node. Besides the high level of availabilitas system, with a large storage so that many learning resources can be gathered by combining learning resources from any educational institution who joined the cloud e-learning by integrating the learning resources with integrated database system mechanism; (4) Just in time learning, using cloud computing for e-learning system encourages the use of e-learning more dynamic with added services through mobile devices, of course, by adding an integrated mobile learning services in a cloud-based e-learning. With adding mobile learning features, cloud-based e-learning become more powerful so the users could access the learning material any-time and any-where and just utilize their mobile devices like smartphones as an example.

7 Conclusion and Future Work

This paper proposes a model of e-learning based on cloud computing. Implementation of e-learning is now generally constructed separately by each institution, the implementation of such this conventional models is costly, because it takes the cost for provision of infrastructure, systems development, and hiring IT staff to maintain and

enhance e-learning systems. Cloud computing as one of the technologies used currently rife in the IT world can be utilized for the implementation of e-learning. With the implementation of e-learning in a cloud environment, educational institutions no longer have to pay for the provision of infrastructure because infrastructure has been provided by the cloud service provider of e-learning and agencies that wish to use it only pay according to the usage by the institution. For the cost of developing e-learning systems and staff to maintain and enhance e-learning systems, cloud service providers also provide service for it, and educational institutions will only pay for the services they already use.

In this paper we propose the architecture that consists of five layers, namely : (1) infrastructure layer; (2) platform layer; (3) application layer; (4) access layer; and (5) user layer. The first three layers are the basic of cloud services, then we added two additional layers, namely access layer and user layer. Access layer consists of a variety of devices used to access the cloud e-learning, whether in the form of notebooks, PCs, Smartphones, Tablets, etc. At the users layer consists of various educational institutions that will use cloud e-learning. As the implementation of the proposed architecture, authors have developed a prototype of a cloud-based e-learning is being piloted at three higher education institutions that are used in teaching and learning, and in the future we will perform an evaluation of the use of cloud-based e-learning.

References

1. Tzeng, G.H., Chiang, C.H., Li, C.W.: Evaluating intertwined effects in e-learning programs: A novel hybrid MCDM model based on factor analysis and DEMATEL. Expert Systems with Applications 32(4), 1028–1044 (2007)
2. Chuang, S., Chang, K., Sung, T.: The Cost Effective Structure For Designing Hybrid Cloud Based Enterprise E-Learning Platform. In: IEEE CCIS (2011)
3. Dong, B., et al.: An E-learning Ecosystem Based on Cloud Computing Infrastructure. IEEE (2009)
4. Chandran, D., Kempegowda, S.: Hybrid E-Learning Platform Based On Cloud Architecture Model: A Proposal. IEEE (2010)
5. Phankokkruad, M.: Implement of Cloud Computing For E-Learning System. In: IEEE ICCIS (2012)
6. Yan, S., et al.: Infrastucture Management Of Hybrid Cloud For Enterprise Users. IEEE (2011)
7. Ghazizadeh, A.: Cloud Computing Benefits And Architecture In E-Learning. IEEE (2012)
8. Carroll, M., Merwe, A., Kotzé, P.: Secure Cloud Computing Benefits, Risks and Controls. IEEE (2011)
9. Pocatilu, P.: Cloud Computing Benefits for E-learning Solutions, Romania, Bucharest. Academy of Economic Studies (2010)
10. Méndez, J.A., González, E.J.: Implementing Motivational Features in Reactive Blended Learning: Application to an Introductory Control Engineering Course. IEEE (2011)
11. Masud, M.A.H., Huang, X.: An E-learning System Architecture based on Cloud Computing. IEEE (2012)
12. Jones, M.T.: Cloud computing and storage with OpenStack: Discover the benefits of using the open source OpenStack IaaS cloud platform. In: Developer Works (2012)

13. Vishwakarma, A.K., Narayanan, A.E.: E-learning as a Service: A New Era for Academic Cloud Approach. In: ISI International Conference on Recent Advances in Information Technology. IEEE (2012)
14. Nuh, M.: Arahan Mendikbud pada Rembug Nasional Pendidikan dan Kebudayaan (2012) (retrieved)
15. Sailah, I.: Kebijakan Direktorat Pendidikan Tinggi Tentang Program Studi. (2012) (retrieved), http://telaga.cs.ui.ac.id/~heru/archives/sarasehanAptikom2
16. Gierlowski, Nowicki: Loosely-Tied Distributed Architecture for Highly Scalable E-Learning System. In: Soomro, S. (ed.) E-learning Experiences and Future. InTech. (2010) ISBN: 978-953-307-092-6
17. Grewal, A., Rai, S., Phillips, R., Fung, C.C.: The E-Learning Lifecycle and its Services: The Web Services Approach. In: Proceedings of the Second International Conference on e-Learning for Knowledge-Based Society, vol. 8, pp. 4.1-4.8 (2005)
18. Pankratius, Sandel, O., Stucky, W.: Retrieving Content With Agents In Web Service e-Learning Systems. In: Symposium on Professional Practice in AI, First IFIP Conference on Artificial Intelligence Applications and Innovations (ALAI), pp. 91–100 (2004)
19. Xu, Z., Yin, Z.G., Saddik, A.E.: A Web Services Oriented Framework for Dynamic E-Leaming Systems. In: IEEE CCECE- CCGEI 2003, Montreal (2003)
20. Arthana, I.K.: Multi-channel Access pada Sistem Temu Kembali Multimedia. Fasilkom UI, Depok (2011)
21. Mell, P., Grance, T.: The NIST Definition of Cloud Computing. NIST, Gaithersburg (2011)

Information Systems Strategic Planning for a Naval Hospital

Hery Harjono Muljo and Bens Pardamean

Bina Nusantara University
Jl. Kebon Jeruk Raya 27, Jakarta 11530, Indonesia
bpardamean@binus.edu

Abstract. This article discusses the Information System (IS) strategic planning for a naval hospital in Indonesia. Its purpose is to improve competitive advantage among hospitals through the addition of new services and products that would lead to improvements in the current patient services. The merging of Hospital Information System (HIS), Radiology Information System (RIS), and Laboratory Information System (LIS) into a single network with a concept of telemedicine is the main topic of this article. The hospital's website is also developed with medical tourism in mind, which attracts more patients, generating more revenue for the hospital.

Keywords: strategic planning, hospital information systems, telemedicine, medical tourism.

1 Introduction

For a company that is planning to upgrade and improve its IS/IT infrastructure, Henderson and Sifonis [1] states it should perform an Information System Strategic Planning (ISSP). The IS Strategic Planning is a useful management tool for identifying the main objective, focus, and work needed for the upgrade [2].

ISSP could also be used as a reference by any organization in designing its IS to perform current business strategics [3] as well as creating new business strategy [4] and IT architectural policy [5]. A well-designed strategy would lead to a competitive advantage [6] while a controlled strategy, which is a critical component in the design, would satisfy both the customer needs and the investor demands [7].

A naval hospital in Indonesia handles not only patients from the Navy but also from the general population that contributes more than 350,000 patients per annum, all of whom are covered by government insurance. The hospital itself has had an IS in place, maintaining the effectiveness and efficiency of day-to-day operations. However, the current system is not equipped for improving service levels or gaining more revenue. Upgrading the system to include these practices place the hospital on par with other competitors that have identified these potentials.

The research question that the author attempts to answer is how would the IS strategic planning be able to perform the following: support the hospital's business

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 202–213, 2013.

strategy, provide a competitive edge, improve current services, and eventually, become the backbone of the hospital's day-to-day operations?

The objectives of the research are to develop an IS strategic planning by analyzing the internal and external side of the hospital's business and the Information Technology (IT) infrastructure, and defining the IT/IS management strategy. The benefit that the hospital could gain from the development is having IT/IS strategies with competitive advantages among the competitors, supporting the hospital's management during the implementation of the IT/IS strategies, and creating new services that would strengthen the hospital as a whole.

2 Methodology

The development of the ISSP refers to the IS/IT strategy development model by Ward and Peppard [8] that consists of five stages. The first stage is preparation and data collection. The preparation sub-stage consists of defining the objectives, limitations, and the problem statement. The second stage is analyzing both the internal and external side of the business environment as well as both the internal and external side of the IT/IS infrastructure.

The tools for analyzing the hospital's external business environment were PEST (Politic, Economic, Social, and Technology), Porter's Five Forces, and Boston Consulting Group (BCG). As for the internal business environment, Value Chain was used to identify the hospital's strengths and weaknesses. Once all the opportunities, threats, strengths, and weaknesses were identified, a SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis was performed to identify the hospital's "position" to achieve the goals. The next step was to find the Critical Success Factors (CFS) of the hospital to find necessary elements required by the hospital for achieving the goals. The third step of the analysis was performing a Gap analysis. The fourth was identifying the business IS strategy, IS/IT management strategy, and the IT strategy. The last stage was planning the IT/IS implementation strategy that would be the blueprint for the hospital.

In this study, a qualitative descriptive analysis approach was utilized for describing the actual hospital condition/environment, identifying the problems, collecting data, analyzing data and situation, as well as finding the answers to the research questions in IS/IT strategic planning form.

The study used surveys, observations, and interviews as primary sources for data analysis. The naval hospital in central Jakarta and two other hospitals were selected as primary data sources. Secondary sources were also collected, such as the population data, inflation rate, gross domestic incomes, and gross domestic expenditures from the Badan Pusat Statistik (Central Bureau of Statistics) website.

3 Result and Discussions

3.1 External Business Environment Analysis

This analysis was used to identify both the external factors related to the organization and the organization's competitive advantages against the competitor. The analytical tools were PEST, Porter's Five Forces, and BCG.

PEST (Politic, Economic, Social, and Technology) Analysis

PEST is a useful tool for summarizing the external factors of a business. PEST prepares a business for handling external factors such as politic, economic, social, and technology. PEST factors tests predictive models for current and future scenarios [9]. PEST analysis would also be used to test the current practice and develop a new framework [10].

Politic. Indonesian Law No. 4 of 2009 Section 6 states that the hospital financing for the poor will incur additional costs for the hospital. Section 7 of the same law states that a hospital has to be established at a geographically strategic location with sufficient treatments, services, and logistics.

Economy. High inflation rate and crude oil prices (more than a $100/barrel in 2012) resulted in sharp increase in drug prices and medical equipment. These conditions place the hospital at an unfavourable disadvantage.

Social. A dense population around the hospital, an increase in revenue and welfare, and various lifestyles are affecting the public health issues of the population.

Technology. Rapid growth in technology and technological literacy population has forced the naval hospital to periodically update its technological infrastructure as often as possible. Competitors that frequently update their technology could be a threat for the naval hospital's management.

Porter's Five Forces

1. Rivalry among Existing Competitors

There are four regular hospitals within the naval hospital area's geographical scope that can be considered as competitors: PELNI Hospital, PK St. Carolus Hospital, PGI Cikini Hospital, and Jakarta Hospital. These competitors are experienced and constantly improving their services, upgrading their medical equipments, renovating their faciities, and upgrading their IS.

2. Threat of New Parties

Sahid Sahirman Memorial Hospital (SSMH) is a new party to watch in the healthcare business competition. SSMH is equipped with modern technology, excellent services, big investment support, and a state-of-the-art IT/IS infrastructure.

3. Threat of Products or Services Substitution

Healthcare products and services substitution such as small local clinics, accupunture clinics, reflexology clinic must also be considered. It is highly likely that patients at one point, will opt to try alternative treatments for their illnesses due to several reasons, such as healthcare cost and treatment effectiveness.

4. Bargaining Power of Buyer

The main customers of the naval hospital are navy officers and the general population. The naval hospital has a strong bargaining power because of the requirement that navy officers check-in to this hospital when they are on duty. The naval hospital also offers a hyperbaric therapy that not many hospitals offer.

5. Bargaining Power of Supplier

The large number of suppliers that exist in the healthcare industry is also a big advantage for the naval hospital because they do not have to rely solely on one party

for its hospital supplies. Thus, the hospital is able to negotiate its payment to the suppliers; the negotiation period can last between one to three months.

Boston Consulting Group (BCG)

BCG Matrix is a trusted and effective technique in product market decisions [11]. The matrix is also used to formulate concept-based restructuring of portfolio analysis and is useful for strategic decision making [12].

According to the BCG Matrix as shown in Figure 1, "outpatient" and "radiology" are both in the "Question Marks" category, which means that both have a low market share but a high market growth. For "inpatient", "laboratory", and "Hyperbaric" are in the "Dogs" category due to their slow market growth and low market share. "Surgical procedure" falls under the "cash cow" category due to its low market growth but high market share. The naval hospital could support all other services through the high profit and high cash generated from the "surgical procedure" service. With low market growth and no recurring investment needed, the naval hospital could keep its profit high. All the cash surplus from the "surgical procedure" service can be used to transform the "Question Marks" category to a "Share Leader" category. It can also be used to subsidize the "Dogs" category, cover the administration fees, fund research and development, as well as cover its debt.

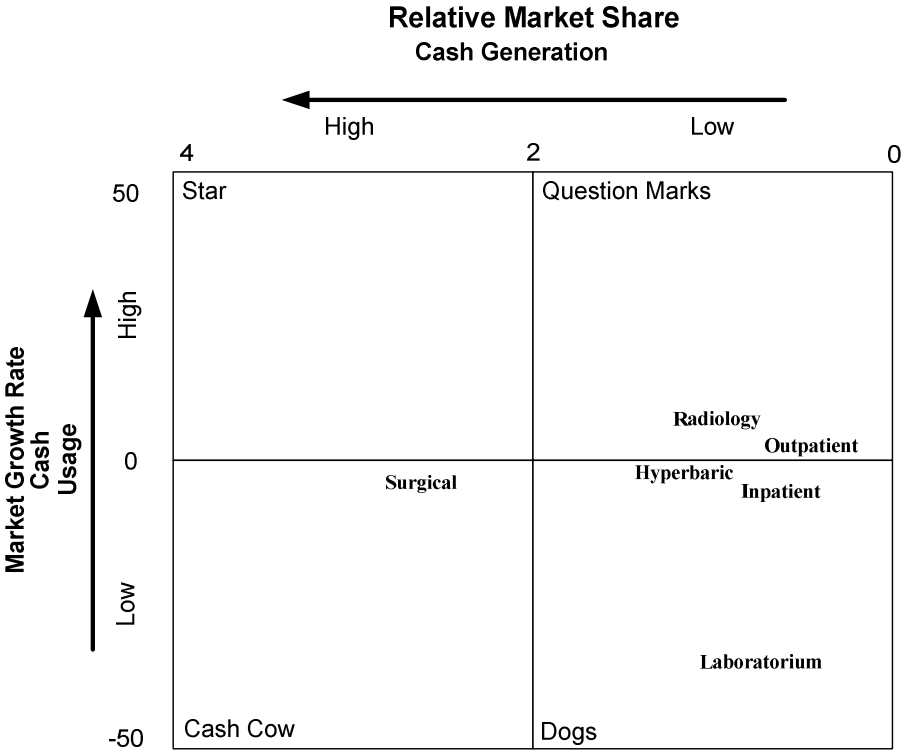


Fig. 1. BCG Matrix

3.2 Internal Business Environment Analysis

This analysis is used to identify the strengths and weaknesses of an organization. Value Chain, SWOT, and Critical Success Factors (CSF) are all the tools that were used to analyze the situation.

Value Chain

Value Chain model by Martinelly, Riane, and Guinet [13] is specifically developed for hospital environment. The main activities from the model include Admissions Logistics, Care Services, Discharge Logistics, Marketing and Sales. Other supporting activities include Infrastructure, Human Resource Management, Technological Development, and Procurement. Figure 2 shows the Value Chain model for the naval hospital.

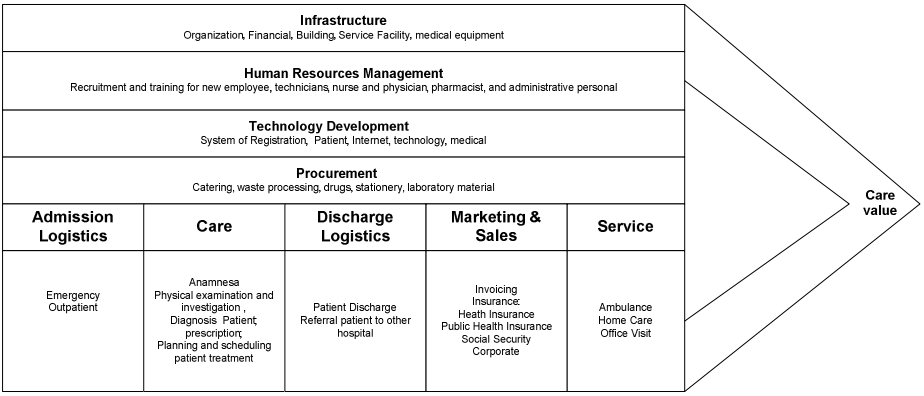


Fig. 2. Value Chain Model

SWOT Analysis

SWOT analysis is a tool that can be used to identify the internal and external strengths and weaknesses of a company [14]. SWOT analysis can help the organization in developing a strategy as well as in preparing for a better future [15]. In an industry, SWOT analysis is used to provide the leader/management team with basic information regarding the company's development policies for future growth [16]. Figure 3 shows the result of SWOT analysis.

Critical Success Factors (CFS)

Critical Success Factors are critical activities needed to obtain a successful (or failing) company. CSF can help provide a better understanding on the needs of customers, as well as better construct the marketing division's strategy, management, and evaluation [17]. The result of the CSF analysis for the naval hospital is shown in Table 1.

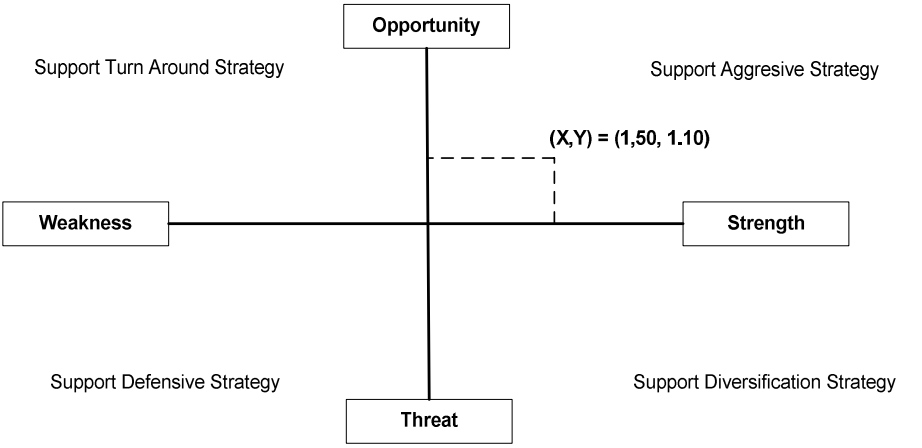


Fig. 3. SWOT Analysis

Table 1. CSF Analysis for the Naval Hospital

Objective	Critical Success Factor	Indicator
To attract out-of-state patient (outside DKI Jakarta)	The number of out-of-state patient (outside DKI Jakarta)	Out-of-state Patient → 10%
To develop an IT-based data processing	The number of IT-based service	IT-based service level → 50%
To have joint agreements with other parties	Joint agreement with other parties (educational institution, investors, competitors, suppliers, and hospitals)	The number of joint agreements in the current year is larger than the previous year's
To increase patient satisfaction	Increasing patient satisfaction	The number of complaints from patients in the current year is fewer than the previous year's

3.3 External IT/IS Environment Analysis

Progress made in the Hospital Information System (HIS), such as Radiology Information System (RIS) and Laboratory Information System (LIS) significantly impacts the hospital's operation. LIS creates a dynamic link among the laboratory analyzers, medical technologists, and clinical providers. LIS, combined with an automated process, has the potential to increase work efficiency and quality, reduce human errors, and lab sample tracking (a task often hampered by human errors) [18]. Additionally, developments in telemedicice and SMS gateway can also increase the competitive field within the healthcare industry. For the current available services, the naval hospital lacks two aspects, the Radiology Information System (RIS) and

Laboratory Information System (LIS). On the other hand, both the naval hospital and its competitors lack the version of HIS that integrates telemedicine, SMS gateway, and medical tourism. Table 2 below lists a comparison between the naval hospital and its competitors in terms of their service applications as well as recommended services that should be developed in order to improve treatment quality.

Table 2. Service Applications Comparison

Application	PK St. Carolus Hospital	PELNI Hospital	Naval Hospital	
			Current	Under Development
HIS				
• Stand Alone	✓	✓	✓	✓
• Integrated	X	X	X	✓
RIS	X	✓	X	✓
LIS	✓	✓	X	✓
Telemedicine Service	X	X	X	✓
SMS Gateway Service	X	X	X	✓
Medical Tourism	X	X	X	✓

The development of Mobile Clinical Assistant (MCA) C5 from the previous version also tremendously impacts the hospital's operations. The development of MCA C5 was propelled by an increase in access patients graphics, especially at the point of care, an increase in clinic's documentation, patient care, and to improve the satisfaction of service providers [19].

3.4 Internal IT/IS Environment Analysis

The current hospital applications does not cover many hospital operations; it only comprises of outpatient, inpatient, and emergency patient registrations. Additionally, the hardware that the hospital owns is rarely upgraded. The Information System Management (IMS) that is currently utilized is developed with Microsoft Visual Studio .NET 2005 and SQL Server 2005. The architecture is a Web-based and Client-Server combination. All the modules use a Web-based architecture, except for the pharmacy and finance modules, which use the Client-Server architecture. The naval hospital application portfolio can be seen in Table 3.

HIS development is aligned with the business strategy to improve services and data processing. Its development also begins with the development of an integrated hospital application that is then followed by the online module. The newly designed modules consist of Patients Master Index module (contains detailed information o the patients), Appointments module (assists patients in appointment set-ups with the hospital), Queue Management module (links the data between receptionists and schedulers), Registration module (records patient registrations manually, online, or through SMS), Outpatient module (manages outpatient transaction), Inpatient and Labor module (manages inpatient transaction), Medical Record Tracking module

(records and tracks patient medical records), Surgical module (manages surgical patients), Emergency module (manages emergency medical records), Pharmacy module (records and manages prescriptions), Billing module (tracks all billing and payments from patients and/or their affiliated third-party entities), and Logistic module (tracks logistical data).

Table 3. The Naval Hospital Applications Portfolio

Strategic	***High Potential***
	1. HIS 2. RIS 3. LIS 4. Telemedicne 5. SMS gateway 6. Medical tourism 7. Home care, hospice
Key Operational	***Support***
Registration Information Systems: 1. In-patient Application 2. Out-patient Application 3. ER Application	Microsoft Office

Following the HIS, Radiology Information System (RIS) is developed to meet radiology's needs for consistency in the database that manages modality, material and medical equipment, and data delivery. RIS is a repository for reports and patient data, as well as electronic records for patient data that could help facilitate appointments, track patient information, and provide online reports on diagnostic results [20].

LIS helps screen more people by increasing the capacity of a health promotion center, and brings in more revenue to the center [21]. The LIS, combined with a Clinical Information System such as Computerized Physician Order Entry and Electronical Medical Record, can support improvements in health care services [22].

The next two developments are telemedicine and SMS gateway. Telemedicine is a medical application in which information is transferred through an interactive audio and visual media. Telemedicine aims to deliver medical consultations and examinations to remote area, thereby reducing the patient's need to travel [23]. Telemedicine could potentially overcome geographical barriers, improve access to patients, facilitate collaboration between health care providers in patient treatment, leading to reduction in patient mortality [24]. This technology is also useful for setting primary and specialty care targets in the public health care system, facilitating electronic communication within the specialty care's referral system, assisting hospitals with a primary care focus in remote areas, and improving cooperation between specialists and the rest of the medical staff [25].

Through the use of video conferencing, patients are better prepared at counseling and expressing their ailments, which would lead to improvments in patients satisfaction levels [26]. Telemedicine could also improve the patients' well-being, independence, self management, medical knowledge, and their overall health.

A module of Virtual Communities for Healthcare can also be used as a virtual self-help guide by the patients [27].

Short Message Service (SMS) gateway is a platform that provides a mechanism to deliver and receive SMS from mobile devices as another way to improve services. It can also be used by the organization to monitor, control, and manage the organization's assets [28]. In a service industry, SMS gateway has been used to increase customer satisfaction e.g. its usage in the hospitality industry for hotel reservations [29]. Additionally, the SMS gateway can also be used to distribute information and to ensure that the customers make a wilfull complaints and able to be contacted again when needed [30].

The high cost of treatments in the United States, Europe, Australia, and other developed countries calls for an increasing demand for a more affordable service. That demand can be answered by the expansion of medical tourism, which is a sector of the healthcare service with a more affordable price [31]. This can be done if the system is supported by an international standard for hospital management. Hospital website development is one IT/IS strategy that can be implemented to support medical tourism in order to recruit patients worldwide as well as provide prospective patients with hospital information, such as service rates and available facilities.

The next development is a collaborative information system that aims to store all data from external parties. This module require the collaboration of drug suppliers, competitors, referral hospitals, and medical schools.

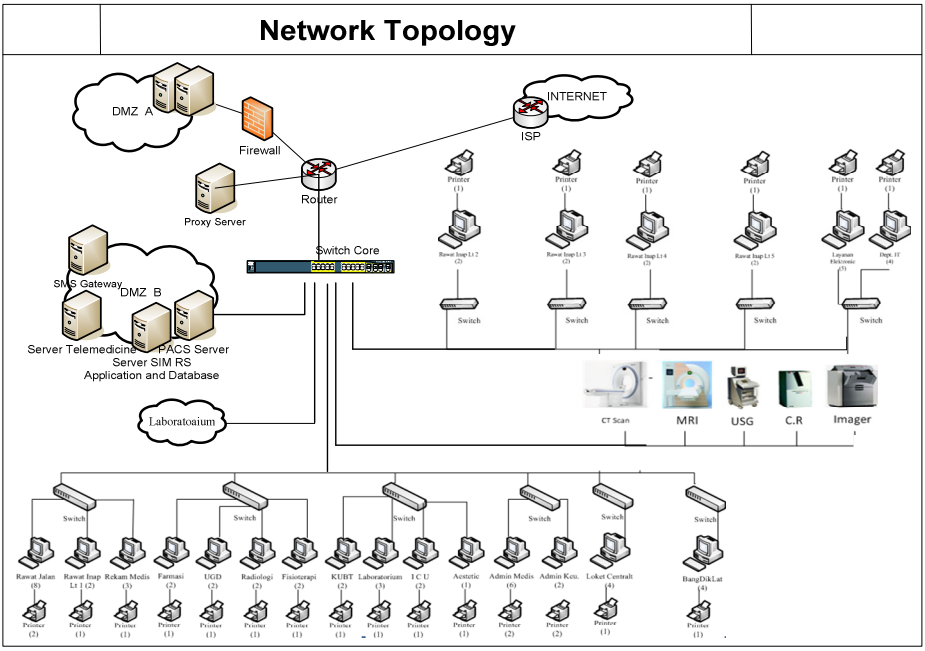


Fig. 4. Overall Strategy IT/IS Network Infrastructure

The last development is an IS patient service that consists of office visits, home care, hospice comfort care, and taxi. Office visit is a patient session with physicians at the doctor's office but not necessarily at the hospital. An appointment module is needed to support this effort. It is a strategy that is developed as a follow-up treatment for post-treatment ambulatory patients. Home care service is provided by the hospital for patients who feel uncomfortable being treated at the hospital. This service could lead to quality enhancement, error reduction, and patient satisfaction improvement [32]. A module needed for this effort is a registration module that can be integrated with the registration module from the HIS. A hospice service is for patients who can no longer be treated and is given comfort care for the "final call". This service can also be integrated with the HIS especially with the registration, billing, and payment modules.

Figure 4 shows the overall network architecture for the proposed IT/IS strategic planning that includes HIS, RIS, LIS, Telemedicine supported, SMS gateway, medical tourism, as well as all other non-medical services.

4 Conclusion

The development of information system strategy provided several added values for the naval hospital's competitive edge. Furthermore, it also redesigned the hospital's applications portfolio. The result of this new strategy offered new products and services such as telemedicine, medical short message service gateway, medical tourism, and home care.

References

1. Henderson, J.C., Sifonis, J.G.: The Value of Strategic IS Planning: Understanding, Consistency, Validity, and IS Markets. MIS Quarterly 12, 186–200 (1988)
2. Adams, J.: Successful Strategic Planning: Creating Clarity. Journal of Healthcare Information Management 19, 3 (2005)
3. Hartog, C., Herbert, M.: Opinion Survey of MIS mManagers: Key Issues. MIS Quarterly 12, 4 (1985)
4. Porter, M.E.: Competitive Advantage: Creating and Sustaining Superior Performance. Free Press, New York (1985)
5. Earl, M.J.: Management Strategies for Information Systems Planning. Prentice Hall, Englewood Cliffs (1989)
6. Wagnera, S.M., Johnson, J.L.: Configuring and Managing Strategic Supplier Portfolios. Industrial Marketing Management 33, 717–730 (2004)
7. Davis, A., Olson, E.M.: Critical Competitive Strategy Issues Every Entrepreneur should Consider before Going into Business. Business Horizons 51, 211–221 (2008)
8. Ward, J., Peppard, J.: Strategic Planning for Information System, 2nd edn. John Wiley & Sons (2002)
9. Ha, H.: E-Government in Singapore A SWOT and PEST Analysis. Asia-Pacific Social Science Review 6, 103–130 (2006)
10. Ayo, C.K., Adebiyi, A.A., Tolulope, F.I., Ekong, U.O.: A Framework for e-Commerce Implementation: Nigeria a Case Study. Journal of Internet Banking and Commerce 13, 1–12 (2008)

11. Armstrong, J.S., Brodie, R.J.: Effects of Portfolio Planning Methods on Decision Making: Experimental Results. International Journal of Research in Marketing 11, 73–84 (1994)
12. Santosa, E., Ling, J.T., Djohanputro, B.: Penerapan Analisis Portofolio untuk Pengambilan Keputusan bagi Pemerintah Provinsi DKI Jakarta pada Perusahaan Patungan. Journal of Management and Business Review 7, 1–12 (2010)
13. Martinelly, D.C., Riane, F., Guinet, A.: A Porter-SCOR Modeling Approach for the Hospital Supply Chain. International Journal of Logistics Systems and Management 5, 436–456 (2009)
14. Pesonen, M., Kurttila, M., Kangas, J., Kajanus, M., Heinonen, P.: Assessing the Priorities using a SWOT among Resource Management Strategies at the Finnish Forest and Park Service. Forest Science 47, 534–541 (2001)
15. Chan, X.: A SWOT Study of the Development Strategy of Haier Group as One of the Most Successful Chinese Enterprises. International Journal of Business and Social Science 2, 147–153 (2011)
16. Narayan, P.K.: Fiji's Tourism Industry: A SWOT Analysis. The Journal of Tourism Studies 11, 15–24 (2000)
17. Sahney, S.: Critical Success Factors in Online Retail – An Application of Quality Function Deployment and Interpretive Structural Modeling. International Journal of Business and Information 3, 144–163 (2008)
18. Voegele, C., Tavtigian, S.V., de Silva, D., Cuber, A.S., Thomas, Calvez-Kelm, F.L.: A Laboratory Information Management System (LIMS) for a High Throughput Genetic Platform Aimed at Candidate Gene Mutation Screening. Bioinformatic Applications Note 23, 2504–2506 (2007)
19. Baltimore, Maryland. Clinician Usability Study: Workflow and Clinician Satis-faction Improvement for Physician CPOE and Nursing eMAR using the Motion C5 with Cerner. UMMS White Paper – C5 Mobile Clinical Assistant (2008)
20. Nor, R.M.: Medical Imaging Trends and Implementation: Issues and Challenges for Developing Countries. Journal of Health Informatics in Developing Countries 5, 89–98 (2011)
21. Chae, Y.M., Lim, H.S., Lee, J.H., Bae, M.Y., Kim, G.H.: The Development of an Intelligent Laboratory Information System for a Community Health Promotion Centre. Asia Pac. J. Public Health 14, 64–68 (2002)
22. Harrison, J.P., McDowell, G.M.: The Role of Laboratory Information Systems in Healthcare Quality Improvement. International Journal of Health Care Quality Assurance 21, 679–691 (2008)
23. Fabbrocini, G., De Vita, V., Pastore, F., D'Arco, V., Mazzella, C., Annunziata, M.C., Cacciapuoti, S., Mauriello, M.C., Monfrecola, A.: Teledermatology: From Prevention to Diagnosis of Nonmelanoma and Melanoma Skin Cancer. International Journal of Telemedicine and Applications, 1–5 (2011)
24. Selinger, S.J., Bates, J., Araki, Y., Lear, S.A.: Internet-Based Support for Cardiovascular Disease Management. International Journal of Telemedicine and Applications, 1–9 (2011)
25. Coelho, K.R.: Identifying Telemedicine Services to Improve Access to Specialty Care for the Underserved in the San Francisco Safety Net. International Journal of Telemedicine and Applications, 1–14 (2011)
26. Dansky, K.H., Bowles, K.H., Palmer, L.: How Telehomecare Affects Patients. Caring 18, 10–14 (1999)
27. Chorbev, I., Sotirovska, M., Mihajlov, D.: Virtual Communities for Diabetes Chronic Disease Healthcare. International Journal of Telemedicine and Applications, 1–7 (2011)

28. Hakim, A.R., Bramanto, A., Syahri, R.: Aplikasi Monitoring Suhu Ruangan Berbasis Komputer dan SMS Gateway. Jurnal Informatika Mulawarman 5, 32–38 (2010)
29. Rahayu, S., Zulmansyah, N.: Perancangan Aplikasi Reservasi Kamar untuk Pelanggan Tetap dengan Menggunakan SMS Gateway. SINAPTIKA 3 (2012)
30. Utomo, H.T., Samopa, F., Setiawan, B.: Pengembangan Sistem Pengaduan Kon-sumen Terkait Bisnis Online Berbasis Facebook Open Graph Protocol dan SMS Gateway. Jurnal Teknik ITS 1, 362–367 (2012)
31. Jones, C.A., Keith, L.G.: Medical Tourism and Reproductive Outsoursing: The Dawning of a New Paradigm for Healthcare. International J. of Fertility and Women's Medicine 51, 251–255 (2006)
32. Thomas, C.R.: The Medical Home: Growing Evidence to Support a New Approach to Primary Care. J. of American Board of Family Medicine 21, 427–440 (2008)

Estimation of Precipitable Water Vapor Using an Adaptive Neuro-fuzzy Inference System Technique

Wayan Suparta* and Kemal Maulana Alhasa

Institute of Space Science (ANGKASA), Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor Darul Ehsan, Malaysia
wayan@ukm.my, kemalalhasa@gmail.com

Abstract. Water vapor has an important role in the global climate change development. Because it is essential to human life, many researchers proposed the estimation of atmospheric water vapor values such as for meteorological applications. Lacking of water vapor data in a certain area will a problem in the prediction of current climate change. Here, we reported a novel precipitable water vapor (PWV) estimation using an adaptive neuro-fuzzy inference system (ANFIS) model that has powerful accuracy and higher level. Observation of the surface temperature, barometric pressure and relative humidity from 4 to 10 April 2011 has been used as training and the PWV derived from GPS as a testing of these models. The results showed that the model has demonstrated its ability to learn well in events that are trained to recognize. It has been found a good skill in estimating the PWV value, where strongest correlation was observed for UMSK station (r = 0.95) and the modest correlation was for NTUS station (r = 0.73). In general, the resulting error is very small (less than 5%). Thus, this model approach can be proposed as an alternative method in estimating the value of PWV for the location where the GPS data is inaccessible.

Keywords: PWV, Adaptive neuro-fuzzy inference system, Estimation, Meteorological applications.

1 Introduction

One of the most important factors in meteorology is determining the rate of water vapor in the atmosphere. Water vapor content and its variability have an important role in human activities and environments. It has been balancing the energy in the preservation atmosphere, the weather process and ultimately important for operational weather forecasting. Water vapor comes from the evaporation of sea water, ground water, rivers, swamps, glaciers, snow, and water in the atmosphere in the form of clouds. Therefore, accurate estimation of water vapor circulation is of primary importance for monitoring and prediction of precipitation rates to indicate climate change takes place at small scales as well as at regional scales. Many efforts have been invested to study the water vapor changes, especially from the science

* Corresponding author.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 214–222, 2013.

communities such as meteorology, hydrology and climatology [1]. Atmospheric water vapor content in terms of precipitable water vapor (PWV) was derived from GPS and the surface meteorically data with superior in temporal and spatial resolution [2],[3]. This technique is still had good capability in comparison with the ground-based techniques that are traditionally made through balloon-borne radiosondes and water vapor radiometers (WVRs). Although the GPS networks currently distributed thousands around the world, there is a still lacking of GPS data, especially in the remote locations that could potentially give the extreme effects on the local weather conditions.

One of the tools of soft computing is the adaptive neuro-fuzzy inference system (ANFIS). ANFIS is a class of adaptive networks that are functionally equivalent to fuzzy inference systems. ANFIS represents both the Sugeno and the Tsukamoto fuzzy models. It has emerged as a powerful tool in solving problems in engineering and non engineering. With uses a hybrid-learning algorithm, it has an ability to do adjustments of rules by using learning set of data and allow the rules to adapt. Thus, it is profoundly suitable for control, pattern recognition and forecasting task. Several researchers have been employed the ANFIS techniques due to its computational speed, robustness, and also its ability to handle complex problems of non-linear functions. Among them is estimation of evaporation using Artificial Neural Networks (AAN) and ANFIS techniques [4], which found that the ANFIS models are better than ANN although the difference is small. Estimation of time series on earthquake events with mapping function [5], and estimation of subsurface strata of Earth [6] reported that the interpretation using ANFIS technique will give the promising results with much fewer percentage error.

The main objective of this study is to investigate the potential use of ANFIS model to predict PWV that influenced by meteorological factors. The second is to evaluate the performance of ANFIS model in estimating the water vapor value by comparing the PWV data obtained from the Global Positioning System (GPS).

2 Methodology

2.1 Data Collection

Two types of data were used in proposed ANFIS. There are the surface meteorological data (pressure, temperature and relative humidity) as the original input and PWV data as the target output. The tested data were taken from the Universiti Malaysia Sabah Kota Kinabalu, Malaysia (**UMSK:** 6.03^{o}N, 116.12^{o}E and height of 63.49 m) and Nanyang Technological University, Singapore (**NTUS:** 1.350^{o}N, 103.680^{o}E and height of 75.38 m). Sabah as a main base of this study has an equatorial climate and naturally more affected by two types of surges associated with East Asian Winter Monsoon that much of the rainfall received during this period. For this work, data gathered from 4 to 10 April 2011 is processed. The meteorological data were taken data of air temperature, barometric pressure, and relative humidity. The PWV data were collected using the ground-based GPS receiver, and the surface meteorological data were collected using the Paroscientific MET4A Broadband meteorological sensors. All the data collected were taken at one–minute interval.

2.2 Adaptive Neuro-fuzzy Inference System

Adaptive neuro-fuzzy inference system is a method that combines neural networks and fuzzy inference system. This method using Sugeno inference model or Takagi-Sugeno-Kang (TSK) fuzzy structures interface, which adapt the rules that are used to fix the parameters so that resulting in a minimum error. ANFIS consists of five components, namely input and output database, fuzzy generator, fuzzy inference system and adaptive neuro. To study the estimation of PWV mode, combination between Sugeno model and TSK inference as a fuzzy inference system (FIS) and adaptive neuron were employed. The optimization method is done by using a hybrid learning algorithm.

Suppose there are two inputs x_1, x_2 to the node and one output is y, then there are two rules in the base Sugeno models [7]:

$$\text{Rule 1: if } x_1 = A_1 \text{ and } x_2 = B_1 \text{ then } f_1 = p_1 x + q_1 y + r_1$$

$$\text{Rule 2: if } x_2 = A_2 \text{ and } x_2 = B_2 \text{ then } f_2 = p_2 x + q_2 y + r_2$$

where x_1 and x_2 are the crisp inputs to the node, and A_1 , B_1 , A_2 , B_2 are fuzzy sets, the reasoning mechanism on this Sugeno model is

$$f = \frac{w_1 f_1 + w_2 f_2}{w_1 + w_2} = \overline{w}_1 f_1 + \overline{w}_2 f_2 \tag{1}$$

where w_1 and w_2 are the degree of membership of a fuzzy set, while f_1 and f_2 are linear equations for the output from a first-order of Sugeno inference model.

ANFIS network consist of five layers (see Figure 1) as follows:

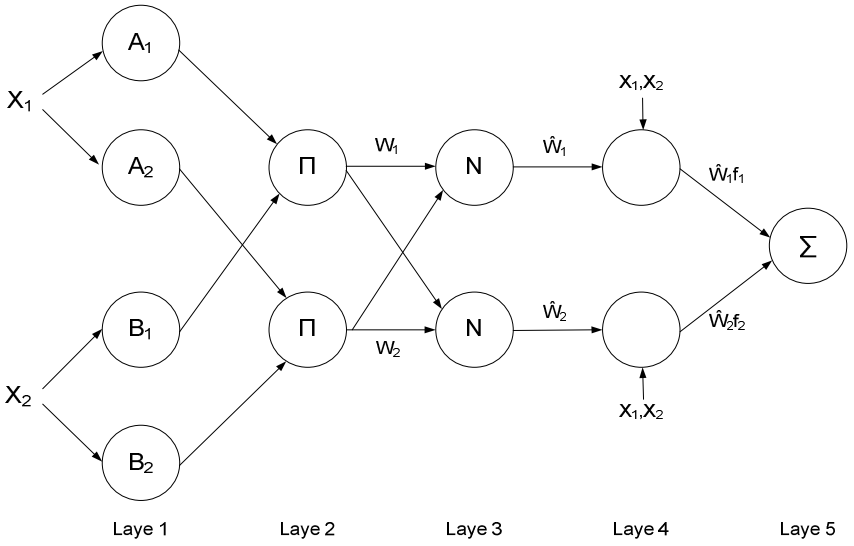


Fig. 1. Architecture of Adaptive Neuro Fuzzy Inference system

Layer 1: This layer serves as fuzzyfication process. The output of node *i* in layer 1 is marked as $O_{l,i}$. Every node *i* in this layer is adaptive to the parameters of an activation function or the output:

$$O_{1,i} = \mu_{Ai}(x), \quad i = 1,2 \tag{2}$$

$$O_{1,i} = \mu_{Bi-2}(y), \ i = 3,4 \tag{3}$$

where x or y is the input value to node *i* and A_i or B_{i-2} is a linguistic label (fuzzy set) associated to the node. While $O_{l,\mathrm{i}}$ is membership degree of fuzzy set. For example, suppose the membership function of fuzzy set A is given as follows:

$$\mu A(x) = \frac{1}{1+\left|\frac{x-c}{a}\right|^{2b}} \tag{4}$$

where {a,b,c} are the parameter set, generally value of b is equal to 1. As the value of these parameters change, the bell curve shape will change as well. The parameter in this layer is usually called the name of the premise parameters.

Layer 2: Every node from this layer is fixed node with labeled π, which output of the result is the incoming signal to entire node. Every output node in this layer present the firing strength of each rule, generally used the AND operator (t-norm operator).

$$O_{2i} = w_i = \mu_{Ai}(x)\mu_{Bi}(y), i = 1,2 \tag{5}$$

For convenience, outputs from this layer generally are called normalized firing strengths.

Layer 3: Every node in this layer is non-adaptive or fixed with denoted by N. Every node serves only to calculate the ratio between the firing strength of i$^{\mathrm{th}}$ rule towards total firing strength of all rules.

$$O_{3i} = \overline{w}_i = \frac{w_i}{w_1+w_2} \tag{6}$$

where $\overline{w}_i$ is normalized firing strengths output of layer 3 and $(p_i x + q_i y + r_i)$ is parameter set on first order Sugeno fuzzy inference system model. The parameters in this layer are referred to as consequent parameters.

Layer 5: A single node in the fifth layer is a fixed node which is the sum of all incoming signals.

$$O_{5i} = \sum_i \overline{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{7}$$

Thus, the fifth layer will build an adaptive network that is functionally equivalent to the first order Sugeno fuzzy inference Model.

3 Result and Discussion

In this study, ANFIS was set with the epoch number of 700, step size 0:01 with a membership function 'gaussmf'. The learning method (adaptive network) used the hybrid algorithm. The data used in this study is the surface meteorological data and the GPS PWV data as mentioned in section 2.1. The data are constructed into two parts, training data and checking data. The data used for training as much 2/3 data and 1/3 the data used for checking.

Before the data input is processed by the ANFIS, the first work is to construct initial FIS. In forming the initial FIS, needed a plan. Among them is the establishment of membership functions and rule formation. The curve used in the formation of the membership function is 'gaussmf'. This is due to the Gaussian curve represent the changing of continuous data. PWV is a continuous, which means its value can change over time depending on the season changes, particularly temperature and air pressure conditions. A Gauss curve using two parameters: the central curve domain and the standard deviation that shows the width of the curve. The two values are determined by using Fuzzy C-Means clustering. The value of the central domain curve is taken from the average value of the data cluster and the standard deviation values were obtained from the standard deviation of a data cluster. The formation rules in Sugeno-Takagi type FIS using a linear equation. There are four linear equations in this FIS that has been optimized to estimate PWV. The reasoning mechanisms on this Sugeno model after training by ANFIS are:

$$f_1 = 0.1309P + 2.567T + 0.6413H - 196.9 \tag{8}$$

$$f_2 = 0.06896P + 2.991T + 0.6319H - 157.3 \tag{9}$$

$$f_3 = -0.1251P + 2.851T + 0.6539H + 39.91 \tag{10}$$

$$f_4 = 0.3626P + 3.031T + 0.8667H - 421.5 \tag{11}$$

where P is the surface pressure (mbar), T is the surface temperature (in degree Celsius), and H is the relative humidity (in percent). If the predicate α for the four rules are W_1, W_2, W_3, and W_4, then the weighted average can be calculated as

$$\mathrm{PWV} = \frac{W_1 f_1 + W_2 f_2 + W_3 f_3 + W_4 f_4}{W_1 + W_2 + W_3 + W_4} = \overline{W_1} f_1 + \overline{W_2} f_2 + \overline{W_3} f_3 + \overline{W_4} f_4 \tag{12}$$

3.1 PWV Result at UMSK

Figures 2a and 2b show the PWV result between ANFIS technique and observed by GPS at UMSK station. In general, the pattern of PWV predicted by ANFIS follows the pattern of PWV observations. Comparing to both figures, data are found blank in Figure 2b due to lack of GPS data at that time and only the surface meteorological data are available and therefore, the time series is contrast to Figure 2a. This is the advantage of the ANFIS model, which use the surface meteorological data as input. ANFIS model was successfully estimating the PWV value and clearly follow the

pattern of PWV observation. Figure 3 shows the scatterplot standard deviation (STD) of PWV between ANFIS and GPS at the UMSK station. The STD calculation was done with one-hour interval of data. The STD pattern between ANFIS and GPS showed a binomial trend where the maximum curve reached to 2.4 mm.

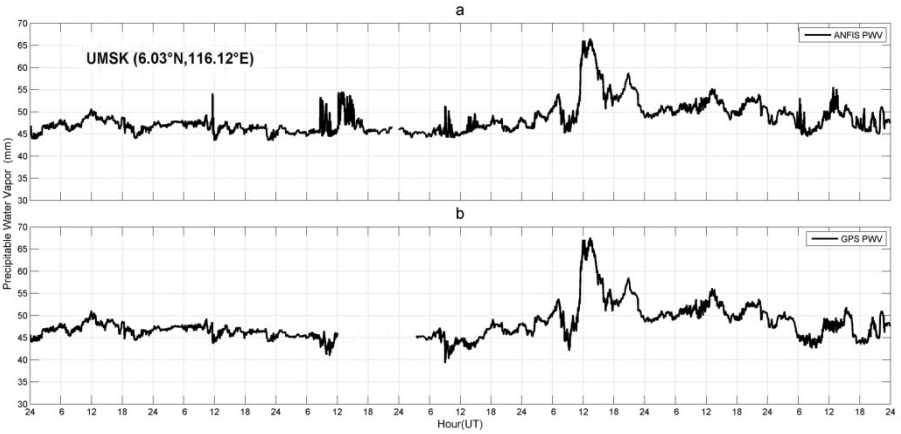


Fig. 2. PWV results between prediction using (a) ANFIS technique and (b) observation using GPS at UMSK station, Malaysia

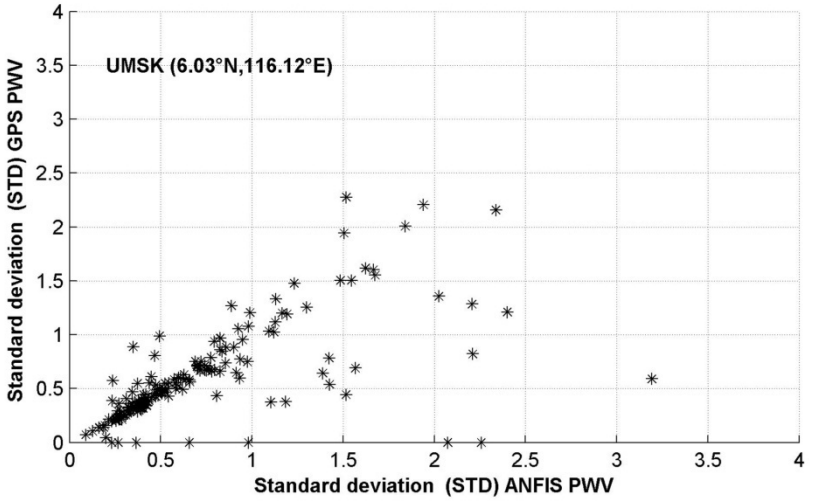


Fig. 3. The scatterplot of STD PWV between ANFIS and GPS for UMSK station

3.2 PWV Result at NTUS

A similar PWV result between ANFIS and GPS was also found for NTUS station as depicted in Figures 4a and 4b. From Figure 4a, the final shape of PWV ANFIS is

clearly similar to the shape of PWV GPS. Looking at the Figure 4b, some data blank for GPS PWV was recorded at NTUS station. This is same case happened like UMSK station, where there is no GPS data provided at this time. From this lacking, ANFIS model successfully solved the problem as demonstrated in Figure 4a. Figure 5 shows the scatterplot of standard deviation (STD) of PWV between ANFIS and GPS. The STD trend at NTUS was found a similar pattern with the UMSK station with a maximum curve reached to 1.82 mm.

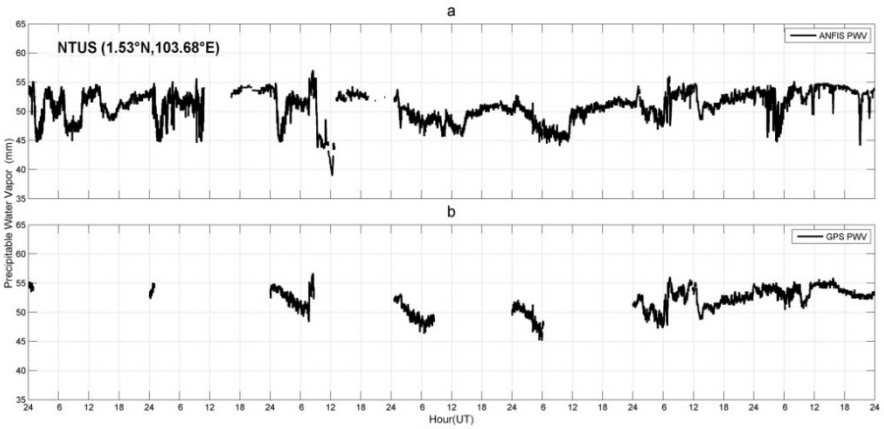


Fig. 4. PWV results between prediction using (a) ANFIS technique and (b) observation using GPS at NTUS station, Singapore

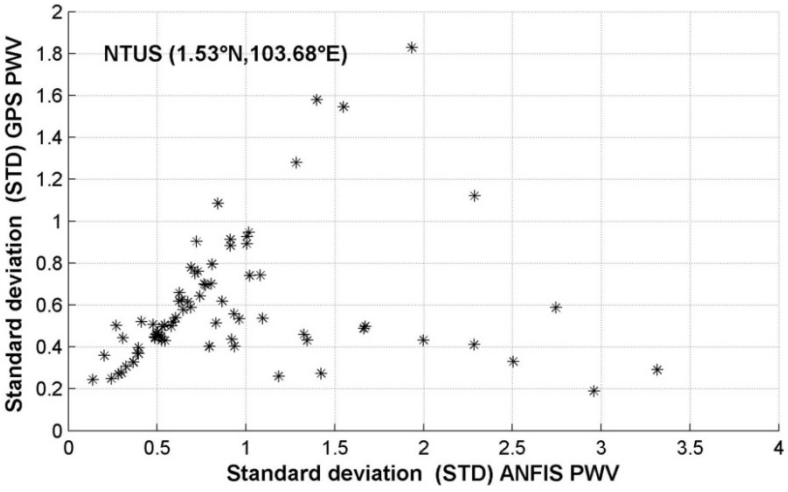


Fig. 5. The scatterplot of STD PWV between ANFIS and GPS for NTUS station

From two cases of examination of ANFIS technique at UMSK and NTUS stations shows that the ANFIS prediction results clearly follow the pattern of PWV observed by GPS in the Southeast Asia region, especially, Malaysia and Singapore. The results of PWV comparison at two example stations are given in Table 1.

Table 1. Statistical parameters of the PWV comparison

Quantity	UMSK	NTUS
R	0.95	0.73
RMSE	1.29 mm	1.68 mm
Mean	-0.06 mm	-1.31 mm

The correlations between the predicted results and observational data show that the ANFIS technique can be used to predict the PWV value, using meteorological data as the input parameter. The relationship with the highest correlation value is observed at UMSK station (r = 0.95), while the modest value is observed at NTUS station (r = 0.73). From calculation of RMSE value, the resulting error is very small which less than 5%. The PWV difference between the prediction error and the observation are obtained -0.06 mm for UMSK station and -1.31 mm for NTUS station. The negative value indicates that the PWV value from ANFIS is lower than those of PWV obtained from GPS. On the other hands, this method is more convenient, cost-effective and practical when compared to the GPS observations. The promising method only needs the meteorological data (pressure, temperature and relative humidity) as input parameters and very helpful for meteorological station with absent of GPS data.

4 Conclusion

In this study, a new method for estimating the PWV value was developed by using one of the tools soft computing, ANFIS techniques. The correlation coefficient between the predicted results and the observational data was found strongest relationship. Looking at UMSK and NTUS stations as an example of examination, the highest correlation coefficient was found at UMSK and the modest correlation coefficient was at NTUS with a resulting error was less than 5%.

A model approaches adaptive neuro fuzzy system interfaces with the input meteorological data (temperature, pressure and relative humidity) can be used as an alternative method in the estimation of PWV value when absent of GPS data in a particular station. In the future, it is recommended to use more extensive data to improve the estimation of PWV value. It is either by adding a new parameter or increases the amount of interval data used to further clarify the accuracy of PWV model developed.

Acknowledgment. This work was partially supported by the Ministry of Higher Education Malaysia (MOHE) under grants UKM-LL-07-FRGS0211-2010 and PKT 1/2003.

References

1. Zhang, S., Xu, L., Ding, J., Liu, H., Deng, X.: Advance in Neural Network Research & Application: A Neural Network Based Algorithm for the Retrieval of Precipitable Water Vapor from MODIS Data. LNEE, vol. 67, pp. 909–916. Springer, Heidelberg (2010)
2. Bevis, M., Businger, S., Herring, T.A., Rocken, C., Anthes, R.A., Ware, R.H.: GPS Meteorology Remote Sensing of Atmospheric Water Vapor Using the Global Positioning System. J. Geophys. Res. 97, 15787–15801 (1992)
3. Suparta, W., Abdul Rashid, Z.A., Mohd Ali, M.A., Yatim, B., Fraser, G.J.: Observation of Antarctic Precipitable Water Vapor and Its Response to The Solar Activity Based on GPS sensing. J. Atmos. Sol.-Terr. Phys. 70, 1419–1447 (2008)
4. Kumar, P., Kumar, D., Jaipaul, Tiwari, A.K.: Evaporation Estimation Using Artificial Neural Networks and Adaptive Neuro-Fuzzy Inference System Techniques. Pakistan Journal of Meteorology 8, 81–88 (2012)
5. Joelianto, E., Widiyantoro, S., Ichsan, M.: Time Series Estimation on Earthquake Events Using ANFIS with Mapping Function. International Journal of AI 3, 37–63 (2009)
6. Srinivas, Y., Stanley, R.A., Hudson, O.D., Muthuraj, D., Chandrasekar, N.: Estimation of Subsurface Strata of Earth Using Adaptive Neuro-Fuzzy Inference System (ANFIS). Acta Geod. Geoph. Hung. 47(1), 78–89 (2011)
7. Jang, J.S.R.: ANFIS: Adaptive Network-Based Fuzzy Inference Systems. IEEE Trans. Syst. Man Cybern. 23, 665–685 (1993)

A Data-Driven Approach toward Building Dynamic Ontology

Dhomas Hatta Fudholi[1,2], Wenny Rahayu[1], Eric Pardede[1], and Hendrik[2]

[1] Department of Computer Science and Computer Engineering, La Trobe University, Australia
dfudholi@students.latrobe.edu.au,
{w.rahayu,e.pardede}@latrobe.edu.au
[2] Department of Informatics, Universitas Islam Indonesia, Indonesia
{hatta.fudholi,hendrik}@fti.uii.ac.id

Abstract. Ontology has been emerged as a powerful way to share common understanding, due to its ability to chain limitless amount of knowledge. In most cases, groups of domain expert design and standardize ontology model. Unfortunately, in some cases, domain experts are not yet available to develop an ontology. In this paper, we extend the possibilities of creating a shareable knowledge conceptualization terminology in uncommon domain knowledge where a standardized ontology developed by groups of experts is not yet available.

Our aim is to capture knowledge and behaviour which is represented by data. We propose a model of automatic data-driven dynamic ontology creation. The created ontology model can be used as a standard to create the whole populated ontology in different remote locations in order to perform data exchange more seamlessly. The dynamic ontology has a feature of a real-time propagation from the change in the data source structure. A novel *delta* script is developed as the base of propagation. In order to complete the model, we also present an information of application support in the form of Jena API mapping for propagation implementation.

Keywords : data-driven, dynamic ontology, propagation.

1 Introduction

Ontology has been used as a mechanism to share common knowledge and understanding [1]. Groups of domain experts have used ontology to represent certain knowledge into semantic structure of information, for instance, in medical health domain. However, there are a large amount of domain knowledge is still untouched by domain experts.

To save time, reduce manual work and facilitate communities who may not have the technical understanding in constructing an ontology, a few researchers have proposed some approaches to develop ontology from underlying data. Garcia et al. and Bohring et al. have done similar research in creating the concept of XML (eXtensible Markup Language) to OWL (Web Ontology Language) mapping, which can be found

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 223–232, 2013.

in [2] and [3]. Both research implement XSD (XML Schema) as the source of creating terminological ontology model. The XSD could be extracted from XML data. XSLT (XML Stylesheet Language Transformation) is used as the tool to translate XML-based information into ontology knowledge representation. Bohring et al. also use XSLT to populate the terminological ontology model. Zhou et al. [4] had research in automatic ontology creation from relational database (RDB). They create seven rules to map the database structure into the terminological conceptualization in ontology, and then populate the records as the ontology instance.

Data source knowledge can change very often. A method to propagate the ontology can be used to keep the ontology dynamic and up-to-date. Sari et al. in [5] propose a propagation model to update sub-ontology of SNOMED CT. This methodology propagates sub-ontology extracted from the main SNOMED CT ontology based on the change log in the SNOMED CT ontology.

Collective knowledge from communities can be extracted to form a formal standard of representation. When it becomes standard, any following knowledge representation could adopt the same terminology. It enables seamless knowledge sharing. The main aim of this research is to create a model for dynamic ontology, derived from a dynamic data source. The dynamic ontology is maintained through a systematic propagation method triggered by changes in the data source structure. The propagation method uses a *delta* script that contains the difference of the previous and the current data structure. When the remote propagation is needed, the use of *delta* script can save the resource rather than sending the whole new data source or the whole new ontology. The novel concept of *delta* script is also proposed in this paper.

The paper is organized as follows. Section 1 is the introduction, capturing the backgrounds, motivations and aims of the research. Section 2 states the related and supporting works for the research. Section 3 elaborates the whole concept model of the data-driven dynamic ontology. Section 4 focuses on the propagation features, starts from the different types of data changes, the *delta* script construction and the propagation process. Section 5 covers the application support for the propagation process in term of *delta* script and programming framework mapping. It also elaborates the case study as to show the implementation.

2 Related Work

The automation of data-driven ontology creation can be very useful for community to share their knowledge in the form of ontology. This also addresses the limitation of technical capability in ontology building. In general, there are two kinds of data sources that are used widely as a data repository, a structured database and semi-structured XML. A number of researchers have explored the techniques to support ontology creation from these two data sources.

Garcia et al. proposes the XSD2OWL. XSD2OWL contains packages based on an XSL (XML Stylesheet Language) that performs a partial mapping from XML Schema to OWL [2]. Even though it consists only of partial mapping that transform XML Schema to OWL, XSD2OWL covers most of ontology semantic structure. The full

mapping table of XSD2OWL is described in [2]. To perform a complete XML-based ontology creation, XSD2OWL cannot be used as a single tool. It needs to be collaborated with XSD extraction tools to extract XSD from its XML data source, e.g. Trang [6] and oXygen XML Editor [7].

Bohring et al. [3] creates a similar mapping concept to translate extracted XSD from XML into OWL ontology. This work explicitly states the way to populate the ontology using the XSLT and the way to perform a mapping of domain and range in ontology properties.

An approach of semi-automatic ontology creation from RDB schema is introduced by Zhou et al. in [4]. The concept is originally created to overcome time consuming and tedious work in creating hand-built ontology. Zhou et al. give an extension in their concept using WordNet to handle similarities in word term. Zhou et al. made seven rules to map the RDB into ontology. All rules can be seen in [4].

Table 1 summarizes works in ontology mapping from XML and RDB. For instance, class or concept in ontology is generated from a *complexType* element in XML and from table or fixed instance value in RDB.

Table 1. General ontology mapping from XML and Relational Database based on Garcia et al. [2], Bohring et al. [3] and Zhou et al. [4] works

Ontology	XML	Relational Database
Class/Concept	*complexType* element	table, fixed instance value
ObjectProperty	*complexType* element	table relation
DatatypeProperty	*simpleType* element, attribute	column
Cardinality (Max, Min)	occurrence (maxOccurs, minOccurs)	constrain (NOT NULL, primary key)
Property Domain and Range	element, XSD datatype	table relation, column, column data type

3 System Design and Concept

The whole system scenario for a data-driven dynamic terminological ontology development can be seen in Fig. 1. The data source is dynamic, shown by the dashed arrow from the old data to the new data. The data source could be an XML data source or RDB data source. The data source consists of data records and data schema. Basically, there are two main parts of the whole concept scenario, the initial ontology creation process and dynamic ontology propagation process.

The initial ontology creation process is depicted using a solid line in Fig. 1. The aim of this process is to create a base dynamic ontology as the very first ontology model to be shared. Schema to ontology translator performs the creation by mapping

the data schema inside the data source into ontology. The dynamic ontology propagation process handles the update of dynamic shared ontology and it will be updated directly when there is a change in the data source.

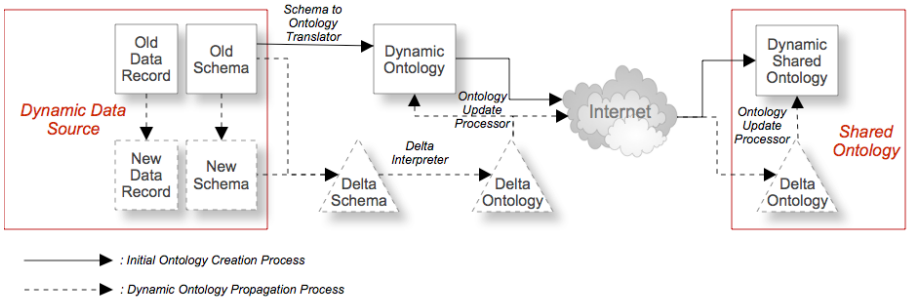


Fig. 1. Conceptual model for creating data-driven dynamic shared ontology to share community knowledge

The dashed lines in Fig. 1. represents the propagation process. The difference of the current and old data schema is stored as *delta schema* script. Since there would be different representation in XML-based schema and RDB schema representation, *delta ontology* is derived from the *delta schema* to create a common representation, which maps directly to the ontology changes. There are two ways of retrieving the current terminological ontology from the shared ontology: (i) by requesting directly the current ontology or (ii) by requesting the *delta ontology* script followed by propagating the ontology locally using propagation application. In addition, there will be only one application needed to use the *delta ontology* script when updating the ontology, since it will be data source type independent. The dynamic ontology propagation process and related tools will be elaborated in Section 4.

4 Dynamic Ontology Propagation

4.1 Changes in Data

Propagation is proposed as the solution to update the common terminological ontology based on the dynamic changes in the data source structure. The structures changes basically consist of *delete*, *insert, rename* and *move*. For XML-based data, all of changes operation could happen in every element and attribute. The *move* change operation of element is the changes in tree structure position of parent and child. RDB's table and column could also have the same change operation; however the *move* operation might be happening only in table column.

4.2 *Delta* Script

The differences of data source structure are gathered in a *delta* script. The purpose of the *delta* script is to patch or upgrade the dynamic ontology. The use of *delta* script

can be useful when the source or the original version is not present in the same location and the ontology needs to be updated without sending the original file, which can be very big. Cobena et al. in [8] give four main benefits for using DIFF (difference) method as change detection: version and querying the past, learning about changes, monitoring changes, and indexing. In addition, Cobena et.al. in [9] proposes about a set of important criterias for a good *delta* script. Those aspects are *Completeness*, *Minimality*, *Performance and Complexity*, *"Move" Operation*, and *Semantics*. All aspects mentioned are considered and applied in the proposed *delta* script.

4.3 *Delta Schema* Script

Delta schema script consists of the difference between the current data schema with the previous version of the schema. It lists all of the difference structure from edit operations. The list includes *delete*, *insert*, *rename*, and *move* list.

Definition DS-1. $\Delta S \equiv \langle D, I, R, M \rangle$. *Delta schema* script comprise of 4 set of list, which are *delete (D)*, *insert (I)*, *rename (R)* and *move (M)*.

The list is proposed to keep up with the *completeness* and *minimality* of the operation. Even though command DELETE and INSERT (we use the all capital words to describe the programming command and to differentiate them from the *delta* script's list and their general common usage words) are the primitive operation and the *delete* and *insert* list could be used to represent *rename* and *move*, but the list will keep the *minimality* aspect and can be directly performed to some programming framework. Therefore, that list can potentially reduce the complexity and yet it is *complete*. The sequence of listed difference in the *delta schema* script should be as mentioned in the **Definition DS-1** to avoid the possible name duplication of the new inserted data and the need to state all inserted and renamed component in order to be the target of moved component. Therefore the sequence should be as follow:

$$\textit{Delete(D)} \rightarrow \textit{Insert(I)} \rightarrow \textit{Rename(R)} \rightarrow \textit{Move(M)}$$

To maintain the semantic information of the data, the following rules need to be applied in *delta schema* script's list:

- **DS-Rule 1** - For all type list: There should be an initial sign to differentiate *complexType* element, *simpleType* element and attribute name in XML, also table and column name in RDB. In XML, sign "(c)" can be used to indicate *complexType* element. As for the attribute, sign "@" can be used as the initial. In RDB, sign "(t)" could be used to indicate table. To simplify the representation of each component, as an example, it could be written as follows:

  <*initial*><*s*><*component name*>

  where <s> is separator sign.

— **DS-Rule 2** - For *insert* list: The information of data type, constrains and location/path (if becomes the child or the part of other component) of the inserted component should be stated clearly. As an example, it could be written as follows:

<initial><s><component name><s><datatype><s><constrain>
<initial><s><component's parent>/<initial><s><new component name>

Since the RDB mapping has some additional information to add when there exist relations in two tables as foreign key. Those relations will create an inverse object property between two concepts. Additional information in the insert list of RDB should be added, such as:

<initial><s><component name> ←→ *<initial><s><new component name>*

— **DS-Rule 3** - For *rename* list: The list should contain the path or location of the renamed component along with the new component name. As an example, it could be written as follows:

<initial><s><component's parent>/<initial><s><component name> →
<initial><s><component's parent>/<initial><s><new component name>

— **DS-Rule 4** - For *move* list: The list should contain the path or location of the moved component along with the new component's parent name. For the new location path, information about data type and constrains need to be incuded to maintain the whole semantic information. As an example, it could be written as follows:

<initial><s><component's parent>/<initial><s><component name> →
<initial><s><component's new parent>/<initial><s><component name><s><datatype><s><constrain>

4.4 *Delta Ontology* Script

Delta ontology script consists of the list of ontology structure change, which is derived from the *delta schema* script based on the mapping in Table 1. There are three types of list in *delta ontology* script; *delete, insert* and *rename* list respectively. The move operation of column in RDB will affect in changing domain and range of property in ontology. Since there is no move operation for domain and range in the ontology, it will trigger the insert and delete operation instead. The move element operation in XML will affect in the ontology restriction. It will not move the restriction to other ontology class but it will trigger a delete and insert operation of the restriction. These two conditions are some reasons why the *move* list is absence in *delta ontology*. The following is the proposed syntax in writing *delta ontology* list:

— Delete List :

- For Class/Concept → *c(name)*
- For ObjectProperty → *op(name)*
- For DatatypeProperty → *dp(name)*

– Insert List :

- For Class/Concept → *c(name, superClass name)*
- For ObjectProperty → *op(name, domain, range, minC**, maxC**)*
- For DatatypeProperty → *dp(name, domain, range, minC**, maxC**)*
- For ObjectProperty domain and range change → *opdr(property name, domain, range, minC**, maxC**)*
- For DataTypeProperty domain and range change → *dpdr(property name, domain, range, minC**, maxC**)*

***minC* and *maxC* is an optional minimum cardinality and maximum cardinality information.

– Rename List :

- For Class/Concept → *c(previous name, current name)*
- For ObjectProperty → *op(previous name, current name)*
- For DatatypeProperty → *dp(previous name, current name)*

When transformed to *delta ontology,* a first character "C", "op" and "dp" is used to stated Class, ObjectProperty and DatatypeProperty respectively. The following example is about the translation process in RDB. The sample from XML is stated along the case study in Section 5.

Example. RDB. There is an additional column created named "author" in "book" table. The data type of "author" is string. The "author" column has NOT NULL constrain. This change could be listed in *delta schema* and *delta ontology* as follows:

DELTA SCHEMA

```
INSERT
  author | string | min-1
  (t)book/author
```

DELTA ONTOLOGY

```
INSERT
  dp(dpauthor, Cbook, string, 1)
```

5 Application Support

To demonstrate the application support, especially in the dynamic ontology propagation concept using *delta* script, the OWL propagation mapping into Apache Jena™ [10] API within a Semantic Web application is developed. The application is used to update OWL ontology from changed data structure. The pattern to apply the Jena API in the application is shown in Fig. 2.

There are three main parts of the Jena programming command block that is applied. **(1)**. Call/open the base OWL ontology model. First, `createOntologyModel()` method is used to create a new ontology model which will be processed in-memory and it is expressed in the default ontology language (OWL). Then `read()` method will call/open OWL file path. **(2)**. Apply and execute Jena API for the propagation. This

part can be filled with any method needed to do the propagation. The types of method are shown in Table 2. **(3)**. Write output of the updated OWL ontology. The `write()` method is used to perform this operation.

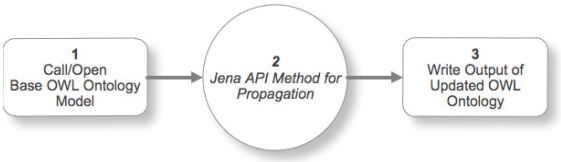


Fig. 2. Jena API Pattern for Propagation Implementation

Table 2. Delta ontology mapping to Jena API

Process	Delta Ontology	Jena API
DELETE	- Class/Concept - DatatypeProperty - ObjectProperty - MinCardinality - MaxCardinality	- `getOntClass()` then `remove()` - `getDatatypeProperty()` then `remove()` - `getObjectProperty()` then `remove()` - `listRestrictions()` then `remove()` - `listRestrictions()` then `remove()`
INSERT	- Class/Concept - DatatypeProperty - ObjectProperty - Set Property Domain - Set Property Range - MinCardinality - MaxCardinality	- `createClass()` - `createDatatypeProperty()` - `createObjectProperty()` - `setDomain()` - `setRange()` - `createMinCardinalityRestriction()` - `createMaxCardinalityRestriction()`
RENAME	Class/Concept or DatatypeProperty or ObjectProperty	`renameResource()`

As a study case, a section of the version 2012 of PubMed/MEDLINE [11] citation XML sample[1] is used. The PubMed/MEDLINE citation XML for the case study and the sample of change in the data can be seen in Table 3. Afterwards, the *delta schema* and *delta ontology* script could be generated as mentioned in Table 4. Fig. 3 depicts a JSP page for the ontology propagation built using Jena API. The input ontology model path, the propagation process step and the output file path can be seen in Fig. 3. Finally, Fig. 4. depicts the Protégé visualization for the propagated ontology based on the change stated in Table 3. It consists of the Class, DatatypeProperty, ObjectProperty and Restriction in the ontology. Due to the limitation of the page, it shows the Restriction for "PubDate" only. From the result, it can be said that the *delta* is complete and holds enough semantic information.

[1] Downloaded from
`http://www.nlm.nih.gov/databases/dtd/medsamp2012.xml`

Table 3. Data sample changes

Previous Data	Current Data
<Journal> <ISSN IssnType="Print">0950-382X</ISSN> <JournalIssue CitedMedium="Print"> <Volume>34</Volume> <Issue>1</Issue> <PubDate> <Year>1999</Year> <Month>Oct</Month> </PubDate> </JournalIssue> <Title>Molecular microbiology</Title> <ISOAbbreviation>M.M./ISOAbbreviation> </Journal>	<Journal> <ISSN IssnType="Print" **CitedMedium="Print"**>0950-382X</ISSN> <JournalIssue> **<Vol>34</Vol>** <Issue>1</Issue> <PubDate> <Year>1999</Year> <Month>Oct</Month> **<Date>4</Date>** </PubDate> </JournalIssue> <Title>Molecular microbiology</Title> </Journal>

Table 4. *Delta* script from sample data changes

Delta Schema	*Delta Ontology*
DELETE ISOAbbreviation INSERT Date | int | min-1 | max-1 (c)PubDate/Date RENAME (c)JournalIssue/Volume → (c)JournalIssue/Vol MOVE (c)JournalIssue/@CitedMedium → (c)ISSN/@CitedMedium | string | min-1 | max-1	DELETE dp(dpISOAbbreviation) INSERT dp(dpDate,CPubDate,int, 1, 1) dpdr(dpCitedMedium,CISSN,string, 1,1) RENAME dp(dpVolume,dpVol)

Propagation

Input Ontology : ...\data\medsamp-part.owl

Output Ontology : ...\data\medsamp-part-output.owl

Delete

- dp(dpISOAbbreviation)

- DatatypeProperty ISOAbbreviation - Deleted

Insert

- dp(dpDate,CPubDate,int, 1, 1)
- dpdr(dpCitedMedium,CISSN,string,1,1)

- DatatypeProperty Date - Created
- DatatypeProperty CitedMedium Domain and Range - Changed

Rename

- dp(dpVolume,dpVol)

- DatatypeProperty Volume - Renamed to Vol

Fig. 3. JSP page for ontology propagation built using Jena API

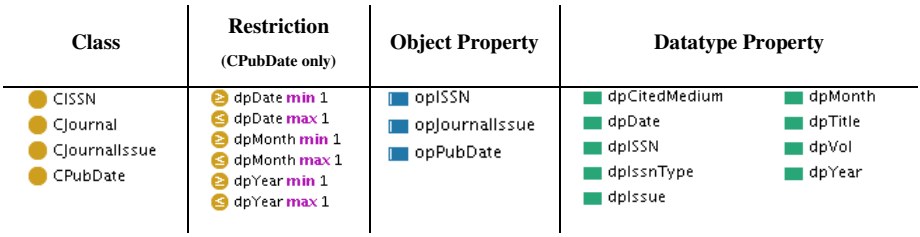


Fig. 4. Protégé visualization of the updated ontology

6 Conclusion

The need to create a common conceptualization from dynamic knowledge has motivated us to create a model for a data-driven dynamic ontology with propagation support. The propagation process updates the base ontology based on the underlying data structure changes. The use of *delta* script gives an advantage in updating remote ontology by sending the minimum source that can provide complete updates. A simple, minimized yet complete *delta* script is designed, and the mapping of the *delta* script list into a Jena API method within a Semantic Web application is demonstrated.

References

1. Calegari, S., Ciucci, D.: Integrating Fuzzy Logic In Ontologies. In: ICEIS (2006)
2. García, R.: A Semantic Web Approach to Digital Rights Management. PhD Thesis. Universitat Pompeu Fabra, Barcelona, Spain (2006)
3. Bohring, H., Auer, S.: Mapping XML to OWL Ontologies. In: Leipziger Informatik Tage. LNI, vol. 72 (2005)
4. Zhou, X., Xu, G., Liu, L.: An Approach for Ontology Construction Based on Relational Database. International Journal of Research and Reviews in Artificial Intelligence 1(1) (2011)
5. Sari, A.K., Rahayu, W., Bhatt, M.: An Approach For Sub-Ontology Evolution In A Distributed Health Care Enterprise. Information Systems Journal (2012)
6. Thai Open Source Software Center Ltd: Trang, Multi-format schema converter based on RELAX NG, http://www.thaiopensource.com/relaxng/trang.html (accessed November 20, 2012)
7. SyncRO Soft SRL, http://oxygenxml.com (accessed November 20, 2012)
8. Cobéna, G., Abiteboul, S., Marian, A.: Detecting Changes in XML Documents. In: Proceedings of the 18th International Conference of Data Engineering (2002)
9. Cobéna, G., Abdessalem, T., Hinnach, Y.: A comparative study of XML diff tools (2004), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.5366
10. The Apache Software Foundation: Apache Jena, http://jena.apache.org/ (accessed November 20, 2012)
11. U.S. National Library of Medicine: MEDLINE®/PubMed® Resources Guide, http://www.nlm.nih.gov/bsd/pmresources.html (accessed November 20, 2012)

Using Semantic Web to Enhance User Understandability for Online Shopping License Agreement

Muhammad Asfand-e-yar and A Min Tjoa

Vienna University of Technology, Favoritenstraße 9-11/188-1, Vienna, Austria
(asfandeyar,amin)@ifs.tuwien.ac.at

Abstract. Normally, a common user sign license agreement without understanding the agreement. License agreements are a form of information, which describes product's usage and its terms and conditions. Habitually, users agree with it but without understanding. In the today's information age, there is no integration of license agreements with any current technology. The contents of license agreements are out of scope for search engines. Management of license agreements using Semantic Web is a multi-disciplinary challenge, involving categorization of common features and structuring the required information in such semantics that is easily extendable and fulfilling the requirements of common user.

In this paper construction of Semantic Web model for Online Shopping license agreement is discussed. The user requirements facilitate the construction of License Ontological model. Moreover, rules are used to capture the complex statements of "terms and conditions". Finally, an explicit semantic model for agreements is constructed that facilitates users' queries.

Keywords: License Agreements, Semantic Web, License Ontological Model, Digital License Agreement.

1 Introduction

In general, installing software or using any services requires the user's agreement to terms and conditions. The terms and conditions describe issues such as functionality, restrictions of use, and about other legal acts such as jurisdiction restriction and intellectual property rights. These terms and conditions describe various legal acts of an agreement. The legal acts contained in agreements are often difficult for a user to understand. Therefore, users do not read such lengthy agreements and bypass the agreement by signing, without understanding. In such situations user might commit illegal act unintentionally. Therefore, an easy understandable solution is required in order to facilitate user's requirements. Hence, a solution in form of agreements' repository is provided that is easy for a user to select an appropriate product. Moreover, Semantic model for online shopping license agreements is discussed in this paper. With the help of the Semantic model user queries are answered in conjunction with user's requirements.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 233–242, 2013.

License Agreements: The agreement is a legal binding between two parties and if any violation occurs then defined penalties are applied. Similarly, using web services requires the user's agreement to the terms and conditions. Agreements describe issues related to product's functionality, restrictions on its use, may specify the number of users that can work in a connected environment, state relevant laws if any, rules pertaining to the distribution of the product, modifications that can be made and the payment procedure [1]. In online shopping centers a legal contract occurs between a buyer and seller while purchasing a product.

The concepts contained in agreements are often difficult for a user to understand. Moreover, the terms and conditions are different nearly in every license agreement. Therefore, user agrees with agreement, without deeper understanding, which might lead to a user's inadvertent implication in an illegal act. Approximately every agreement applies penalties, if terms and conditions are violated.

User Awareness: For the awareness and understanding agreements, few efforts are done for example; GPL-violations, it is an awareness project about the violation of GNU licensing agreements. “GPL-violation” is a communication platform between all parties that are involved in licensing of open source software for example authors and copyright holders, vendors and users [2].

Business Software Alliance (BSA) is an IT industry group that helps their member groups in controlling piracy of software and hardware tools. BSA provides awareness and educate people about the ethical and digital security risk associated with use of unlicensed software [3].

The work done by GPL-violations and BSA serves for user awareness; while the Semantic model provides a solution, in selecting a required license according to the needs of a user, and also make awareness of penalties.

2 Related Work

In the literature, a plethora of studies has been conducted for developing a Semantic Web model to fulfill user requirements. The research is based on two sections, initially, it summaries fundamental concepts of license agreements, by considering different type of existing licenses. Secondly, it highlights the use of ontology in information retrieval, management, automatic extracting of meta-data and different concepts related to Semantic Web. According to the multifaceted nature of the research work, we will describe the problem from more than one angle.

2.1 License Agreements

Organizations do their best to control the misuse of License Agreements. Hence, penalties are defined to assure the legal use, defined in “terms and conditions”.

Various license verifying systems have been developed. These systems help to ensure that a licensed version of a product is properly and legally installed on each computer in an organization. Microsoft verifies the installing products online, before completing the installation of the software product on a system [4]. This clarifies that

the granter of license wants her product to be used legally and according to the terms and conditions defined in the license agreement.

Online shopping agreements are mostly similar to each other and have a lower degree of complexity as software licenses [5]. In these types of agreements, rights of service providers and customers are clearly defined. In some agreements service providers are the sellers and provide a facility of purchasing from Web sites. While other agreements are totally different from service provides, for example in case of eBay. Only two types of online shopping agreements are explained and used as pars pro toto in this paper, i.e. Amazon [6] and eBay [7, 8] license agreements.

2.2 Semantic Web

Digital Right Expression Languages and Policy Management are the two areas, covered in this paper. Digital Right Management is an access control technology that is used to impose limitations on usage of digital contents and devices. It controls the illegal use of digital media by preventing access, copy or conversion to other formats by end users. Semantic Web technologies facilitates in creating ontology for copyright agreements. These copyright agreements are digital rights used to implement legal limitation.

Digital Rights Expression Languages

Rights Expression Languages (REL) reflects licenses legal requirements, by gathering terms and required relations of contents. REL creation is based on an ontology using existing standards for example ODRL, XrML and Creative Commons. Usage of contents depends on the rights of a license provider, granted to licensee. REL with DRM translates licenses and uses ontology to deduce copyrights, context description and different type of expressions used in a license agreement. The REL ontology terms are mapped to a legal dictionary, which consists of legal classifications and legal rules used in the license law [9].

Privacy Policy Matching

Nyre discusses the issues related to "privacy policy matching" [10]. They suggested, arranging policies of service providers according to customer requirements. Customer requirements are translated to policies using privacy preferences of the provider, for matching the customer requirements with the service provider polices. Privacy preference provider is introduced as an intermediate channel between service provider and customers; to solve, the issue of matching and arranging the policies for service providers.

The word "policy" used in this method is similar as terms and conditions of license agreement. These policies are defined by the service provider, for example in online shopping license agreements. The main problem of this method is to maintain an intermediate channel as "privacy preference provider" for updates of customer requirements. The second problem is to translate customer's requirements to policies, at this intermediate level, and then match these policies to service provider's policies.

Handling these problems, according to the above method, is worse in case to deal with complex policies of service providers. To solve the above mentioned problem, we propose in this paper a solution using Semantic Web technologies.

3 Use Case Scenarios

Use case scenario elaborates user's requirements to achieve specific results. License model is constructed to facilitate a common user. Therefore, each license is categorized similarly. The model facilitates a common user to acquire her required information from a license. The model is designed to easily adopt documented agreement using Semantic technologies. The Semantic Web model of agreements covers all common features of an agreement. A user scenario is discussed below;

Online Shopping
Using services of online shopping web sites, the policies should be agreeable by both sellers and buyers. Sellers defined their own polices, without conflicting the police of online shopping. In case of conflicting policies, seller has either to face penalties or in worst case could be ban to use the services. This scenario is explained below:

A seller wants to sell her items through online shopping web site. The used items are placed for biding by the seller. One of the buyers received the item, with functionality problem. According to polices of online shopping, a buyer can claim for reimbursement. Therefore, the buyer claimed for reimbursement. In this case, the seller didn't define a proper procedure for reimbursement. The online shopping company paid for the damages, and ceased the seller's account and banded him from selling her item from the site. Such problems normally happens when user bypass the agreement, before using services.

4 Semantic Web and License Agreements

Semantic Web uses ontology as its basic component and uses taxonomy of information as an expressing knowledge to design ontology. Defining a domain of knowledge for license agreements, a root ontology (named; License Ontology) is created using bottom up approach, as discussed in [5, 11].

The ontological structure for license agreements is categorized in two parts, i.e. License Ontology and sub license ontologies. The License Ontology is developed from common features of different license agreements. The License Ontology is extended to sub license ontologies, in order to design ontology as a meaningful license agreement. Sub-license ontology describes a specific license agreement.

The resulting model is equipped with rules according to license agreements. Functionality of entire process depends on search and comparison of user requirements using License Ontology. Query results obtained from sub license ontologies are sent to user for final approval.

4.1 Sub License Ontology

Sub ontology is extracted from parent ontology to elaborate a section of domain knowledge. The extracted sub-ontology has following two features 1) consistency and 2) completeness. Sub-ontology concept is explained by proposing four methods i.e. extend, add, merge and replace. We used extended method of sub-ontology concept. The extend method depends on new features of sub-ontology engineering, which includes concepts, properties, relationships and mapping. These sub license ontologies are categorized into Online Shopping License Ontology and Software License Ontology. The Software License Ontology is explained in detail in [5].

Online Shopping License Ontology
Construction of Online Shopping License Ontology is based on two different agreements. These include Amazon license agreement and eBay license agreement. These agreements have most common features. Online Shopping License Ontology is an extension of License Ontology, as shown in Fig.1. The Online Shopping License Ontology extends all the features of License Ontology. A class named "Reimbursement" is added to Online Shopping License Ontology. The class is required to provide a concept of reimbursement method for customer, after purchasing product. Required properties are added to define the relationship of reimbursement class with customer and seller classes in the Online Shopping License Ontology.

Online Shopping License Ontology has defined individuals. These defined individuals are connected with each other, according to agreements, to define complete meaning of License Ontology

5 Rules Used for Constructing License Ontology

Rules are needed for inference about classes and properties, mapping ontologies, transforming data in different format, using complex queries, axioms and many more. The aim is to populate the domain of License Ontology by means of rules, created form analyzing set of text documents. The entities are used in License Ontology and sub license ontologies on basis of terms and condition of agreements. Rule applied on Online Shopping License Ontology is described as follows;

Disjoint Classes in Ontology
In License Ontology only those classes that have information related to terms and conditions of agreement are considered disjoint. In License Ontology, classes used under "Agreement Items" class, i.e. "Breach", "Core Agreement" and "Non Core Agreement", and also the sub-classes of these defined classes are disjoint to each other.

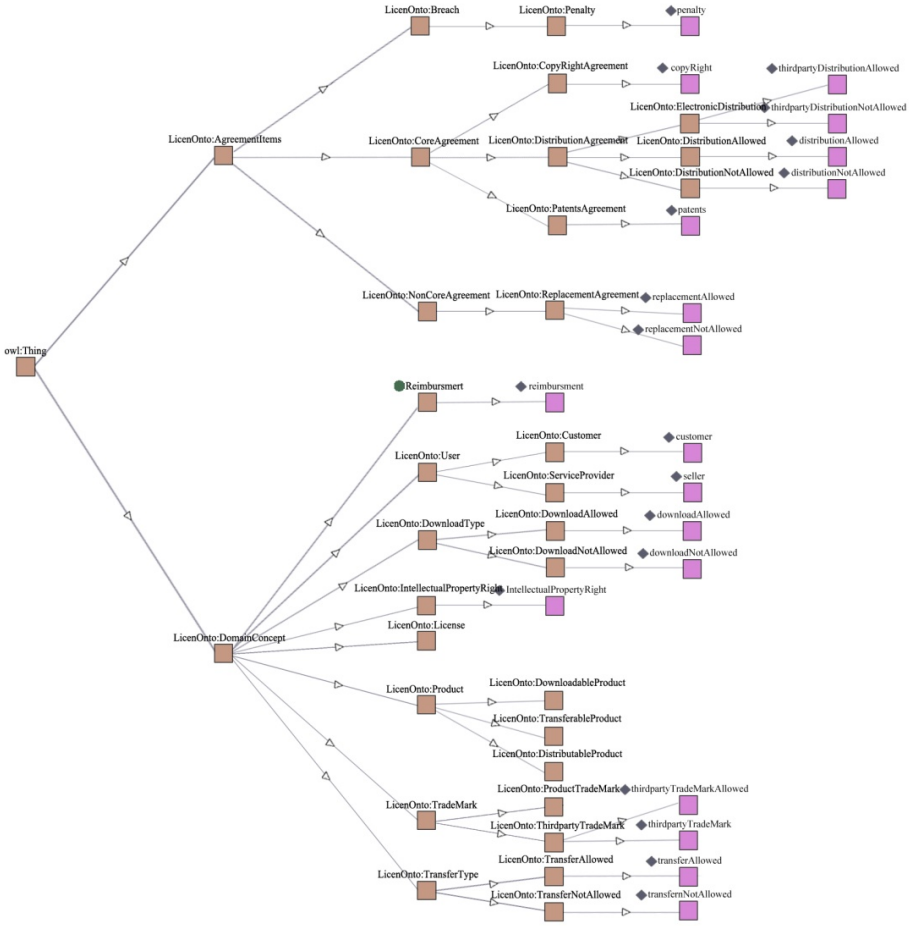


Fig. 1. Online Shopping License Ontology

Restriction Used for Product Class

Details about product's legal usage are documented by license agreement i.e. reflected as below;

$$\text{Product} \sqsubseteq \forall \text{ isDocumentedBy . License}$$

It states that for all values of product should have an agreement (i.e. "License"). Vice verse the restriction applied also on "License" class because both properties are inverse to each other;

$$\text{isDocumented} \equiv \text{documents}^{-}$$

6 Reasoning License Ontology

Acquiring information from ontology can be determined by rules. After having information from ontology by applying rules, then the next step is to represent the information. In this section, ten rules are introduced that are built according to previously discussed user scenarios. These rules describe complete requirements of a user, to search for appropriate license using ontologies.

Online Payment
Rule for online payment provides information, about procedure of payment according to an agreement. The rule describes, that a product have ontology. And according to the ontology, it is inquired that payment procedure supports online payment of the product or not. Then availability of user permissions is checked, for online purchasing of the product, according to the ontology. Thus, according to this rule, an end user is allowed to purchase the product according to an agreement as described below;

(?product LicenOnto:hasLicense ?license) (?license LicenOnto:supportPaymentMethod LicenOnto:Online) (?user profile:paymentMethod profile:Online) ⟶ (?user profile:canPayOnline ?product)

Method of Receiving a Product
The following rule is about downloading a product according to ontology. It states that a product having agreement and allows a user to download the product. The rule examines whether a user can download the product or not. According to ontology, if antecedent part holds answer then consequent part of the rule will also hold answer and allow user to download the product online. The product could be e-book, etc.

(?product LicenOnto:hasA ?license) (?license LicenOnto:receiveMethod SoftOnto:Download) (?user profile:receiveMethod profile:Online) ⟶
(?user SoftOnto:canDownload ?product)

Replacing of Damaged Product
"Replacing of Damaged Product" rule is about replacing a product, if the product is damage before receiving. Therefore, according to ontology the availability of replacing damaged product is evaluated against respective agreement. If a method of replacing a damaged product is allowed according to ontology then the product is allowed for replacement. According to this rule, the possibility of replacing the damaged product is determined according to agreement.

(?user profile:ownsProduct ?product) (?product LicenOnto:hasLicense ?license) (?license LicenOnto:replaceMethod SoftOnto:Return) ⟶
(?user SoftOnto:canReturn ?product)

Reimbursement of Payment
Rule of reimbursement describes the method of reimbursement of expenses after identifying fault in a received product. In agreement normally the reimbursement procedure is defined for customers. Therefore, in ontology a rule is used to inform customer about availability of reimbursement procedure according to an agreement. Whenever, such situation occurs then following rule is used to find out the availability of reimbursement method, according to agreement.

(?user profile:hasPurchased ?product) (?product LicenOnto:hasLicense ?license) (?license LicenOnto:replaceMethod OnliShopOnto:reimbursement) ⟶
(?user OnliShopOnto:canReturn OnliShopOnto:reimbursement)

7 Conclusion

Initially, undertaking an agreement without understanding is a major issue. Habitually customers agree with terms and conditions of an agreement without understanding. Incomplete knowledge and understanding about a license agreement in general and agree with an agreement without understanding specifically - is a main problem for one's own self. To solve the issue, a License Ontology is constructed, which is not only a platform for different license agreements to reside but users can access their required information by querying ontology. The application is designed in such a way that different license agreements can be easily plotted on same platform with minor changes according to agreement.

Based on ontologies, Digital License Agreement application (user interface of License Ontology) uses triples to define a meaningful sentence. Therefore, the results are based on the triples. These triples are defined so that the meaning of a statement in a license agreement should not be changed and the triples should be meaningful according to common user understanding.

8 Future Work

The Semantic Web model for license agreements is extendable, to adopt other license agreements, for example agreements of service providing companies, hardware products, etc. Following agreements could also be model using the License Ontology.

Cross Boundary Agreements and its Conflicts
"Airline's Terms and Condition" could be constructed with similar method, as used for previous explained sub ontologies. Whenever, centralization of same categories of license agreements are achieved, then the ontology will be able to find out the conflicts between different laws (i.e. terms and conditions or license agreements). For example, airlines have to make agreements to use airports and airspace of other territories. Normally, settling downing the conflicts in agreements, consumes maximum time. Such system will not only explain the terms and conditions of each partner but

will also explain the compromising statements that took place between them. These compromising statements will be helpful for upcoming organization to settle down their conflicts with the same/other organization.

SLAs and Business Processes

Cloud computing is a blend of concepts evolved from the SOA, grid computing and Virtualization. Adapting to this new way of computing is a hindrance for the companies to meet their requirements [12]. Laws and policies issues that must be addressed while adopting cloud computing for particular jurisdiction. Some of the challenging issues are; access, reliability, security, liability, IPR, ownership, portability and auditability. The failure of these issues causes resistance to provide services. Lingering mistrust has a negative impact on service providers. Cloud providers also face jurisdictional issues, such as: government intervention and costs of doing business [13].

We propose a Semantic solution for SLA cloud services which supports the consumers in finding the appropriate services that matches their specific requirements. As well as with awareness, consumers have choices to select the proper service according to their requirements. The proposed approach aims at globally unified SLA ontology. Organizations using this approach are believed to be benefited as (i) the categorization of agreements helps to define their "terms and conditions" using Semantic Web technology accordingly (ii) easy in understanding and comparing the different agreements for choosing a better option.

Mapping ontology will be used to relate business process and functions to cloud computing. Semantic rules encapsulate the tacit knowledge of enterprise business. The rules and relations used in Semantic Web technology facilitates businesses in selecting proper service of cloud computing. The selection of services depends on the requirements of customers according to business processes and functions.

References

1. Alliance, B.S.: Why a License matters? (2012), http://ww2.bsa.org/country/Anti-Piracy/WhyaLicenseMatters.aspx
2. Welte, H.: About the gpl-violations.org project (2010), http://www.gpl-violations.org/
3. BSA. About BSA & Members (2012), http://ww2.bsa.org/country/BSAandMembers.aspx
4. Microsoft. Genuine Windows: FAQ (2012), http://windows.microsoft.com/en-US/windows/help/genuine/faq
5. Asfandeyar, M., Anjomshoaa, A., Weippl, E.R., Tjoa, A.M.: Exploiting ontology for software license agreements. International Journal of Software and Informatics (IJSI) 4, 89–100 (2010)
6. Services, A.W. AWS Customer Agreement (2012), http://aws.amazon.com/de/agreement/
7. eBay. Your User Agreement (2012), http://pages.ebay.ie/help/policies/user-agreement.html

8. Post, E.T. License Agreement-eBay Trading Post (2012), http://pages.ebay.ie/tradingassistants/license-agreement.html
9. Nadah, N., Rosnay, M.D., Bachimont, B.: License Digital Content With a Generic Ontology Escaping From The Jungle of Rights Expression Languages. In: International Conference on Hybrid Information Technology, pp. 65–69. ACM (2007)
10. Nyre, A.A., Bernsmed, K., Solvar, B., Pedersen, S.: A Server-Side Approach to Privacy Policy Matching. In: 1st International Workshop on Privacy by design (PDB), ARES 2011 (2011)
11. Asfand-e-yar, M., Anjomshoaa, A., Weippl, E.R., Tjoa, A Min: Blending the Sketched Use Case Scenario with License Agreements Using Semantics. In: Karagiannis, D., Jin, Z. (eds.) KSEM 2009. LNCS, vol. 5914, pp. 275–284. Springer, Heidelberg (2009)
12. Anjomshoaa, A., Tjoa, A Min: How the Cloud Computing Paradigm Could Shape the Future of Enterprise Information Processing. In: Proceedings of 13th International Conference on Information Integration and Web-based Applications Services (iiWAS). IEEE, Ho Chi Minh City (2011)
13. Jaeger Paul, T., Lin, J., Justin, M.G., Shannon, N.S.: Where is the cloud? Geography, economics, environment and jurisdiction in cloud computing. 1st Monday, Peer Reviewed Journal on the Internet 14(5) (2009)

Secure and Verifiable Outsourcing of Sequence Comparisons

Yansheng Feng[1], Hua Ma[1], Xiaofeng Chen[2,⋆], and Hui Zhu[3]

[1] Department of Mathematics, Xidian University,
Xi'an 710071, P.R. China
fengyansheng1108@163.com, ma_hua@126.com
[2] State Key Laboratory of Integrated Service Networks,
Xidian University, Xi'an 710071, P.R. China
xfchen@xidian.edu.cn
[3] Network and Data Security Key Laboratory of Sichuan Province,
Chengdu 611731, P.R. China
zhuhui@xidian.edu.cn

Abstract. With the advent of cloud computing, secure outsourcing techniques of sequence comparisons are becoming increasingly valuable, especially for clients with limited resources. One of the most critical functionalities in data outsourcing is verifiability. However, there is very few secure outsourcing scheme for sequence comparisons that the clients can verify whether the servers honestly execute a protocol or not. In this paper, we tackle the problem by integrating the technique of garbled circuit with homomorphic encryption. As compared to existing schemes, our proposed solution enables clients to efficiently detect the dishonesty of servers. In particular, our construction re-garbles the circuit only for malformed responses and hence is very efficient. Besides, we also present the formal analysis for our proposed construction.

Keywords: Outsourcing, Garbled Circuit, Verifiable Computation.

1 Introduction

Several trends are contributing to a growing desire to outsource computing from a device with constrained resources to a powerful computation server. This requirement would become more urgent in the coming era of cloud computing. Especially, cloud services make this desirable for clients who are unwilling or unable to do the works. Although attractive these new services are, concerns about security have prevented clients from storing their private data on the cloud. Aiming to address this problem, secure outsourcing techniques are developed.

Among the existing secure outsourcing techniques, a specific type receives great attention, namely sequence comparisons. Atallah et al. first propose this concept in [1], which achieves the nice property of allowing resource-constrained devices to enjoy the resources of powerful remote servers without revealing their

⋆ Corresponding author.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 243–252, 2013.

private inputs and outputs. Subsequent works [2, 3] are devoted to obtain efficiency improvements of such protocols. Techniques for securely computing the edit distance have been studied in [5, 6], which partition the overall computation into multiple sub-circuits to achieve the same goal. Besides, [7, 8] introduce the Smith-Waterman sequence comparisons algorithm for enhancing data privacy. Blanton et al. [4] utilize finite automata and develop techniques for secure outsourcing of oblivious evaluation of finite automata without leaking any information. In particular, considering highly sensitive individual DNA information, it is indispensable to privately process these data, for example, [17] encrypts sensitive data before outsourcing. Furthermore, when the length of sequences is large, it is not surprising to alleviate clients from laborious tasks by outsourcing related computation to servers.

Recently, Blanton et al. [11] propose a non-interactive protocol for sequence comparisons, namely, a client obtains the edit path by transmitting the computation to two servers. However, the scheme [11] is impractical to some extent in that it does not achieve verifiability. Also, the scheme [11] has to garble each circuit when the sub-circuit is processed and hence is not efficient. As we know, there seems few available techniques for secure outsourcing of sequence comparisons, which can provide verifiability and enjoy desirable efficiency simultaneously. Among, [15] also achieves the verifiability that can be done by providing fake input labels and checking whether the output from the servers matches a precomputed value, but it verifies with some probability by using the technologies of commitment and Merkle hash tree that differ from our scheme.

Our Contribution. In this paper, we propose a construction for secure and verifiable outsourcing of sequence comparisons, which enables clients to efficiently detect the dishonesty of servers by integrating the technique of garbled circuit with homomorphic encryption. Specially, our solution is efficient in that it regarbles the circuit only for malformed responses returned by servers. Formal analysis shows that the proposed construction is proved to achieve all the desired security notions.

2 Preliminaries

2.1 Edit Distance

We briefly review the basic algorithm for the edit distance [1]. Let $M(i, j)$, for $0 \leq i \leq m$, $0 \leq j \leq n$, be the minimum cost of transforming the prefix of λ of length j into the prefix of μ of length i. So, $M(0, 0) = 0$, $M(0, j) = \sum_{k=1}^{j} D(\lambda_k)$ for $1 \leq j \leq n$ and $M(i, 0) = \sum_{k=1}^{i} I(\mu_k)$ for $1 \leq i \leq m$. Furthermore, we can recurse to obtain the results:

$$M(i,j) = min \begin{cases} M(i-1, j-1) + S(\lambda_j, \mu_i) \\ M(i-1, j) + I(\mu_i) \\ M(i, j-1) + D(\lambda_j) \end{cases}$$

where $S(\lambda_j, \mu_i)$ shows the cost of replacing λ_j with μ_i, $D(\lambda_j)$ shows the cost of deleting λ_j, and $I(\mu_i)$ shows the cost of inserting μ_i.

2.2 Grid Graph Measures

The relevances among the entries of the M-matrix induce an $(m+1) \times (n+1)$ grid directed acyclic graph (DAG). It is apparent to see that the string editing problem can be regarded as a shortest-path problem on DAG. An $l_1 \times l_2$ DAG is a directed acyclic graph whose vertices are the $l_1 l_2$ points of an $l_1 \times l_2$ grid, and such that the only edges from point (i, j) are to points $(i, j+1)$, $(i+1, j)$, and $(i+1, j+1)$. Figure 1 shows an example of DAG and explains that the point (i, j) is at the ith row from the top and the jth column from the left. As special cases of the above definition, the meanings of $M(0, 0)$ and $M(m, n)$ are easy to be obtained.

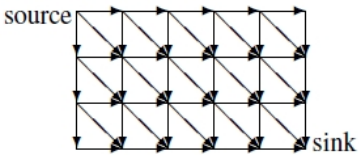


Fig. 1. Example of a 3×5 grid DAG

Such edit scripts that transform λ into μ are in one-to-one correspondence to the edit paths of G that start at the source (which represents $M(0, 0)$) and end at the sink (which represents $M(m, n)$).

2.3 Yao's Garbled Circuit

We summarize Yao's protocol for two-party computation [9], which is initiated by introducing the millionaire problem that is the same as [16]. For more details, we refer to Lindell and Pinkas' description [12].

We assume two parties, A and B, wish to compute a function F over their private inputs a and b. First, A converts F into a circuit C. A garbles the circuit and obtains $G(C)$, and sends it to B, along with garbled input $G(a)$. A and B then engage in a series of OTs so that B obtains $G(b)$ with A learning nothing about b. B then applies the garbled circuit with two garbled inputs to obtain a garbled version of the output: $G(F(a, b))$. A then maps this into the actual output.

In more detail, A constructs the garbled circuit as follows. For each wire w in the circuit, A chooses two random values $k_w^0, k_w^1 \xleftarrow{R} \{0,1\}^\lambda$ to represent 0/1 on that wire. Once A has determined every wire value, then forms a garbled version of each gate g (See Fig.2). Let g be a gate with input wires w_a and

w_a	w_b	w_z
k_a^0	k_b^0	$E_{k_a^0}(E_{k_b^0}(k_z^{g(0,0)}))$
k_a^0	k_b^1	$E_{k_a^0}(E_{k_b^1}(k_z^{g(0,1)}))$
k_a^1	k_b^0	$E_{k_a^1}(E_{k_b^0}(k_z^{g(1,0)}))$
k_a^1	k_b^1	$E_{k_a^1}(E_{k_b^1}(k_z^{g(1,1)}))$

Fig. 2. Circuit's garbled table

w_b, and output wire w_z. Then the garbled version $G(g)$ consists of simply four ciphertexts:

$$\gamma_{ij} = E_{k_a^i}(E_{k_b^j}(k_z^{g(i,j)})), where\ i \in \{0,1\}, j \in \{0,1\} \quad (1)$$

When k_a^α, k_b^β, and the four values γ_{ij} are given, it is possible to compute $k_z^{g(\alpha,\beta)}$. By then B transmits $k_z^{g(\alpha,\beta)}$ to A, A can map them back to 0/1 values and hence obtains the output of the function.

3 Secure Model and Definitions

In our work we transfer the problem of the edit distance computation by a client C for strings $\mu_1, ..., \mu_m$ and $\lambda_1, ..., \lambda_n$ to two computational servers S_1 and S_2. Furthermore, the security requirement is such that neither S_1 nor S_2 learns anything about the client's inputs or outputs except the lengths of the input strings and the alphabet size. More formally, we assume that S_1 and S_2 they are semi−honest and non−colluding, they follow the computation but might attempt to learn extra information. Here, we assume that S_2's ability is more powerful than S_1, we only prove that the attempt of S_2's attack fails in latter proof. Since the powerful adversary can not attack client successfully, neither can the weak one. Security in this case is guaranteed if both S_1's and S_2's views can be simulated by a simulator with no access to either inputs or outputs other than the basic parameters, and such simulation is indistinguishable from the real protocol. We introduce several definitions in the following:

Definition 1. *We say that a private encryption scheme (E, D) is Yao-secure if the following properties are satisfied*:

- Indistinguishable encryptions for multiple messages
- Elusive range
- Efficiently verifiable range

Definition 2. (Correctness). *A verifiable computation scheme VC_{edit} is correct if for any choice of function F, the key generation algorithm produces keys $(PK, SK) \leftarrow Keygen(F, \lambda)$ such that, $\forall x \in Domain(F)$, if $ProbGen_{SK}(x) \rightarrow \sigma_x$, $Compute_{PK}(\sigma_x) \rightarrow \sigma_y$, then $y = F(x) \leftarrow Verify_{SK}(\sigma_y)$.*

We then describe its correctness by an experiment:

Experiment $\mathrm{Exp}_A^{veri}[VC_{edit}, F, \lambda]$
$(PK, SK) \leftarrow Keygen(F, \lambda)$;
For $i = 1, ..., l = poly(\lambda)$; $poly(\cdot)$ is a polynomial.
$x_i \leftarrow A(PK, x_1, \sigma_1, ..., x_i, \sigma_i)$;
$(\sigma_i) \leftarrow ProbGen_{SK}(x_i)$;
$(i, \hat{\sigma}_y) \leftarrow A(PK, x_1, \sigma_1, ..., x_l, \sigma_l)$;
$\hat{y} \leftarrow Verify_{SK}(\hat{\sigma}_y)$
If $\hat{y} \neq \perp$ and $\hat{y} \neq F(x_i)$, output '1', else '0';

The adversary succeeds if it produces an output that convinces the verification algorithm to accept on the wrong output for a given input.

Definition 3. (Security). *For a verifiable computation scheme VC_{edit}, we define the advantage of an adversary A in the experiment above as*:

$$Adv_A^{Verif}(VC_{edit}, F, \lambda) = Prob[Exp_A^{verif}[VC_{edit}, F, \lambda] = 1] \tag{2}$$

A verifiable computation scheme VC_{edit} is secure, if

$$Adv_A^{Verif}(VC_{edit}, F, \lambda) \leq negli(\lambda) \tag{3}$$

where $negli(\cdot)$ is a negligible function of its input. The F in the above descriptions is the function to calculate the $\theta(\cdot)$ in our protocol.

Definition 4. (One-time secure). *It is the same as Definition* 3 *except that in experiment Exp_A^{Verif}, the adversary can query the oracle $ProbGen_{SK}(\cdot)$ only once and must cheat on that input. VC_{Yao} is a special case of VC[1] when the input is single.*

Similar to formulas (2)(3), we can obtain:

$$Adv_A^{Verif}(VC_{Yao}, F, \lambda) = Prob[Exp_A^{verif}[VC_{Yao}, F, \lambda] = 1] \tag{4}$$

$$Adv_A^{Verif}(VC_{Yao}, F, \lambda) \leq negli(\lambda) \tag{5}$$

4 Secure and Verifiable Outsourcing of Sequence Comparisons

4.1 High-Level Description

To gain the edit path, we can use a recursive solution: In the first round, instead of computing all elements of M, we compute the elements in the "top half" and the "bottom half" of the matrix respectively. Then calculate each $M(m/2, j)$ and determine the position with the minimum sum from top to bottom. In [11]

[1] This verifiable computation VC can be consulted in [14].

this position is expressed as $M(m/2, \theta(m/2))$. Then, we discard cells from the top right and lower left of $M(m/2, \theta(m/2))$. We recursively apply this algorithm to the remaining of the matrix. But this case exposes $M(m/2, \theta(m/2))$ to the servers. With protecting $\theta(m/2)$, we form two sub-problems of size 1/2 and 1/4 of the original [11].

We now introduce how this computation can be securely outsourced. First, client produces garbled circuit's random labels corresponding to its inputs (two labels per input bit). Then it sends all the labels to S_1 for forming the circuit and one label per wire that corresponds to its input value to S_2. Once the circuit is formed, S_2 will evaluate it using the labels. The novelty of this way is, scheme without OT protocols is also feasible.

An advanced non-interactive protocol has been proposed in [11]. Based on this paper, we achieve the multi-round inputs verifications by integrating the garbled circuit with fully homomorphic encryption [10]. Specifically, the client will encrypt labels under a fully homomorphic encryption public key. A new public key is generated for each-round input in order to prevent being reused. The server can then evaluate labels and send them to the client, who decrypts them and obtains $F(x)$. This scheme can reuse the garbled circuit until the client receives a malformed response, which is more efficient than generating the new circuit every time.

4.2 The Proposed Construction

The following is a safe and verifiable protocol to achieve the calculation of the edit path. Among, C stands for the client, S_1, S_2 as two servers. m, n are the lengths of two private strings. This allows us to obtain the overall protocol as follows:

Input: C has two private sequences as well as their cost tables, C must generate $min(m, n)$ key pairs of the fully-homomorphic encryption against all the sub-circuits. Besides, using generated keys, C encrypts the private input label's values and delivers these into the circuit.
Output: C obtains the edit path. S_1 and S_2 learn nothing.
Protocol VC_{edit}:

1. **Pre-computing**: C generates two random labels (l_0^t, l_1^t) for each bit of its input $\mu_1, ..., \mu_m, \lambda_1, ..., \lambda_n, I(\mu_i)$ for each $i \in [1, m], D(\lambda_j)$ and $S(\lambda_j, \cdot)$ for each $j \in [1, n], I(\cdot), D(\cdot)$, and $S(\cdot, \cdot)$, $t \in [1, S_\Sigma(m+n) + S_C(m+2n+n\sigma)]$. C also generates $min(m, n)$ key pairs of the fully-homomorphic encryption and runs the encryptions, $(\sigma_x \leftarrow Encrpyt(PK_{\mathbb{E}}^i, l_{bt}^t),\ i \in [1, min(m, n)])$ against all the sub-circuits.
2. **Circuit's construction**: C sends all (l_0^t, l_1^t) to S_1, S_1 uses the pairs of labels it received from C as the input labels in constructing a garbled circuit that produces $\theta(m/2)$, strings $\mu_1^{'}, ..., \mu_{m/2}^{'}, \lambda_1^{'}, ..., \lambda_n^{'}$ and the corresponding $I(\mu_i^{'}), D(\lambda_j^{'})$, and $S(\lambda_j^{'}, \cdot)$, as well as strings $\mu_1^{''}, ..., \mu_{m/2}^{''}, \lambda_1^{''}, ..., \lambda_{n/2}^{''}$ and the relevant $I(\mu_i^{''}), D(\lambda_j^{''})$, and $S(\lambda_j^{''}, \cdot)$. Let the pairs of the output labels that

correspond to $\theta(m/2)$ be denoted by $(\hat{l_0^t}, \hat{l_1^t})$, where $t \in [1, \lceil \log(n) \rceil]$, the labels corresponding to the output labels for the first sub-problem be denoted by $(l_0^{'t}, l_1^{'t})$, where $t \in [1, S_\Sigma(m/2+n) + S_C(m/2+n+n\sigma)]$ and the labels corresponding to the output labels for the second sub-problem be denoted by $(l_0^{''t}, l_1^{''t})$, where $t \in [1, S_\Sigma(m/2+n/2) + S_C(m+n+n\sigma)/2]$. Then, S_1 stores three types of labels.

3. **Keygen**: S_1 transfers $(\hat{l_0^t}, \hat{l_1^t})$ to the client as private key SK. In the pre-computing above, C has already generated key pairs of the fully-homomorphic encryption, $((PK_{\mathbb{E}}, SK_{\mathbb{E}}) \leftarrow Keygen(\lambda))$. C stores SK and $SK_{\mathbb{E}}$ and exposes $PK_{\mathbb{E}}$ to S_2.
4. **Evaluation**: C transmits all the labels σ_x in the pre-computing to S_2 for storing and evaluation. Then, S_2 uses $PK_{\mathbb{E}}$ to calculate $Encrypt(PK_{\mathbb{E}}, \gamma_i)$. Next, runs $Evaluate(C, Encrypt(PK_{\mathbb{E}}, \gamma_i), Encrypt(PK_{\mathbb{E}}, l_{bt}^t))$ with homomorphic encryption's property, we can obtain the $\sigma_y \leftarrow Encrypt(PK_{\mathbb{E}}, \hat{l_{bt}^t})$, which is stored in S_2 later.
5. **Sub-circuits evaluation**: S_1 and S_2 now engage in the second round of the computation, where for the first circuit S_1 uses pairs $(l_0^{'t}, l_1^{'t})$ as the input wire labels as well as the pairs of the input wire labels from C that correspond to cost tables $I(\cdot)$, $D(\cdot)$, and $S(\cdot,\cdot)$. After the circuit is formed, S_1 sends to S_2 who uses the encrypted labels stored before to evaluate this circuit. S_2 saves every result value of the evaluation as $\sigma_y^{(i)}$.
6. **Verification**: When S_1 and S_2 reach the bottom of recursion, S_2 sends the all $\sigma_y^{(i)}$ from each circuit to C. C uses $SK_{\mathbb{E}}$ stored before to decrypt $\sigma_y^{(i)}$ to get $\hat{l_{bt}^t}$. Further, converts the output labels into the output of the function (e.g., $F(x) = \theta(a), a = 1, ..., m$) by using SK, from which it can reconstruct the edit path.

5 Analysis of the Proposed Construction

5.1 Robust Verifiability

In practice, the view of the client will change after the evaluation. How to deal with this situation that client receives a malformed response? One option is to ask the server to run the computation again. But this repeated request informs the server that its response was malformed, server might generate forgeries. The client aborts after detecting a malformed response, but it can hinder our protocol's execution. We will consider it as follow:

There is indeed an attack if the client does not abort. Specifically, the adversary can learn the input labels one bit at a time by XOR operation [13]. So, it can generate cheating. When a malformed response come to the client, this paper needs to continue running the protocols VC_{edit} instead of terminating. we must ask the client to regarble the circuit, every time when a malformed response is received.

5.2 Security Analysis

Theorem 1. *Let E be a Yao-secure symmetric encryption scheme and $\mathbb{E}$ be a semantically secure homomorphic encryption scheme. Then protocol VC_{edit} is a secure and verifiable computation scheme.*

Proof: Since E is a Yao-secure symmetric encryption scheme, then VC_{Yao} is a one-time secure verifiable computation scheme (Proof of Theorem 3 in [13]). Our method is transformming (via a simulation) a successful adversary against the verifiable computation scheme VC_{edit} into an attacker for the one-time secure protocol VC_{Yao}. Next, for the sake of contradiction, let us assume that there is an adversary A such that $Adv_A^{Verif}(VC_{edit}, F, \lambda) \geq \varepsilon$, where ε is non-negligible in λ. We use A to build another adversary $A^{'}$ which queries the ProbGen oracle only once, and for which $Adv_{A'}^{Verif}(VC_{Yao}, F, \lambda) \geq \varepsilon^{'}$, where $\varepsilon^{'}$ is close to ε. Once we prove the Lemma 1 below, we have our contradiction and the proof of Theorem 1 is complete.

Lemma 1. *$Adv_{A'}^{Verif}(VC_{Yao}, F, \lambda) \geq \varepsilon^{'}$ where $\varepsilon^{'}$ is non-negligible in λ.*

Proof: This proof proceeds by defining a set of experiments.
$\mathcal{H}_A^k(VC_{edit}, F, \lambda)$: For $k = 0, ..., l-1$. Let l be an upper bound on the number of queries that A makes to its ProbGen oracle. Let i be a random index between 1 and l. In this experiment, we change the way the ProbGen oracle computes its answers. For the jth query:

- $j \leq k$ and $j \neq i$: The oracle will respond by (1) choosing a random key pair for the homomorphic encryption scheme $(PK_{\mathbb{E}}^j, SK_{\mathbb{E}}^j)$ and (2) encrypting random λ-bit strings under $PK_{\mathbb{E}}^j$.
- $j > k$ or $j = i$: The oracle will (1) generate a random key pair $(PK_{\mathbb{E}}^i, SK_{\mathbb{E}}^i)$ for the homomorphic encryption scheme and (2) encrypt σ_x (label by label) under $PK_{\mathbb{E}}^i$.

We denote with $Adv_A^k(VC_{edit}, F, \lambda)$= Prob$[\mathcal{H}_A^k(VC_{edit}, F, \lambda)$=1].

- $\mathcal{H}_A^0(VC_{edit}, F, \lambda)$ is identical to the experiment $Exp_A^{Verif}[VC_{Yao}, F, \lambda]$. Since the index i is selected at random between 1 and l, we have that

$$Adv_A^0(VC_{edit}, F, \lambda) = \frac{Adv_A^{Verif}(VC_{edit}, F, \lambda)}{l} \geq \frac{\varepsilon}{l} \tag{6}$$

- $\mathcal{H}_A^{l-1}(VC_{edit}, F, \lambda)$ equals the simulation conducted by A' above, so

$$Adv_A^{l-1}(VC_{edit}, F, \lambda) = Adv_{A'}^{Verif}(VC_{Yao}, F, \lambda) \tag{7}$$

If we prove $\mathcal{H}_A^k(VC_{edit}, F, \lambda)$ and $\mathcal{H}_A^{k-1}(VC_{edit}, F, \lambda)$ are computationally indistinguishable, that is for every A

$$\mid Adv_A^k[VC_{edit}, F, \lambda] - Adv_A^{k-1}[VC_{edit}, F, \lambda] \mid \leq negli(\lambda) \tag{8}$$

if we are done above, then that implies that

$$Adv_{A'}^{Verif}(VC_{Yao}, F, \lambda) \geq \frac{\varepsilon}{l} - l \cdot negli(\lambda) \tag{9}$$

The right of inequality is the desired non-negligible ε'. □

Remark: Eq.(8) follows from the security of the homomorphic encryption scheme. The reduction of the security of $\mathbb{E}$ with respect to Yao's garbled circuits to the basic security of $\mathbb{E}$ is trivial. For more details, please refer to [13].

6 Conclusions

This work treats the problem of secure outsourcing of sequence comparisons by a computationally limited client to two servers. To be specific, the client obtains the edit path of transforming a string of some length into another. We achieve this by integrating the techniques of garbled circuit and homomorphic encryption. In the proposed scheme, client can detect the dishonesty of servers according to a response returned from those servers. In particular, our construction re-garbles the circuit only when a malformed response comes from servers and hence is efficient. Also, the proposed construction is proved to be secure in the given security model.

Acknowledgements. We are grateful to the anonymous referees for their invaluable suggestions. This work is supported by the National Natural Science Foundation of China (Nos. 60970144 and 61272455), the Nature Science Basic Research Plan in Shaanxi Province of China (No. 2011JQ8042), and China 111 Project (No. B08038).

References

1. Atallah, M., Kerschbaum, F., Du, W.: Secure and private sequence comparisons. In: ACM Workshop on the Privacy in Electronic Society, WPES (2003)
2. Atallah, M., Li, J.: Secure outsourcing of sequence comparisons. In: Workshop on Privacy Enhancing Technologies, PET, pp. 63–78 (2004)
3. Atallah, M., Li, J.: Secure outsourcing of sequence comparisons. International Journal of Information Security 4(4), 277–287 (2005)
4. Blanton, M., Aliasgari, M.: Secure outsourcing of DNA searching via finite automata. In: DBSec, pp. 49–64 (2010)
5. Huang, Y., Evans, D., Katz, J., Malka, L.: Faster secure two-party computation using garbled circuits. In: USENIX Security Symposium (2011)
6. Jha, S., Kruger, L., Shmatikov, V.: Toward practical privacy for genomic computation. In: IEEE Symposium on Security and Privacy, pp. 216–230 (2008)
7. Kolesnikov, V., Sadeghi, A.-R., Schneider, T.: Improved Garbled Circuit Building Blocks and Applications to Auctions and Computing Minima. In: Garay, J.A., Miyaji, A., Otsuka, A. (eds.) CANS 2009. LNCS, vol. 5888, pp. 1–20. Springer, Heidelberg (2009)

8. Szajda, D., Pohl, M., Owen, J., Lawson, B.: Toward a practical data privacy scheme for a distributed implementation of the Smith-Waterman genome sequence comparison algorithm. In: Network and Distributed System Security Symposium, NDSS (2006)
9. Yao, A.: How to generate and exchange secrets. In: IEEE Symposium on Foundations of Computer Science, FOCS, pp. 162–167 (1986)
10. Gentry, C.: Fully homomorphic encryption using ideal lattices. In: Proceedings of the ACM Symposium on the Theory of Computing, STOC (2009)
11. Blanton, M., Atallah, M.J., Frikken, K.B., Malluhi, Q.: Secure and Efficient Outsourcing of Sequence Comparisons. In: Foresti, S., Yung, M., Martinelli, F. (eds.) ESORICS 2012. LNCS, vol. 7459, pp. 505–522. Springer, Heidelberg (2012)
12. Lindell, Y., Pinkas, B.: A proof of Yao's protocol for secure two-party computation. Journal of Cryptology 22(2), 161–188 (2009)
13. Gennaro, R., Gentry, C., Parno, B.: Non-interactive Verifiable Computing: Outsourcing Computation to Untrusted Workers. In: Rabin, T. (ed.) CRYPTO 2010. LNCS, vol. 6223, pp. 465–482. Springer, Heidelberg (2010)
14. Parno, B., Raykova, M., Vaikuntanathan, V.: How to Delegate and Verify in Public: Verifiable Computation from Attribute-Based Encryption. In: Cramer, R. (ed.) TCC 2012. LNCS, vol. 7194, pp. 422–439. Springer, Heidelberg (2012)
15. Blanton, M., Zhang, Y., Frikken, K.B.: Secure and Verifiable Outsourcing of Large-Scale Biometric Computations. In: IEEE International Conference on Information Privacy, Security, Risk and Trust, PASSAT, pp. 1185–1191 (2011)
16. Vivek, S.S., Selvi, S.S.D., Venkatesan, R., Rangan, C.P.: A Special Purpose Signature Scheme for Secure Computation of Traffic in a Distributed Network. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, JoWUA 3(4), 46–60 (2012)
17. Wang, J., Ma, H., Tang, Q., Li, J., Zhu, H., Ma, S., Chen, X.: A New Effcient Veriable Fuzzy Keyword Search Scheme. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, JoWUA 3(4), 61–71 (2012)

Syntactic Analysis for Monitoring Personal Information Leakage on Social Network Services: A Case Study on Twitter

Dongjin Choi[1], Ilsun You[2], and Pankoo Kim[1,*]

[1] Dept. of Computer Engineering Chosun University,
375 Seoseok-dong, Dong-gu, Gwangju, Republic of Korea
Dongjin.Choi84@gmail.com, pkkim@chosun.ac.kr
[2] Korean Bible University,
16 Danghyun 2-gil, Nowon-gu, Seoul, Republic of Korea
isyou@bible.ac.kr

Abstract. Social network services such as Twitter and Facebook can be considered as a new media different from the typical media group. The information on social media spread much faster than any other traditional news media due to the fact that people can upload information with no constrain to time or location. Because of this reason, people got fascinated by SNS and it sinks into our life. People express their emotional status to let others know what they feel about information or events. However, there is a high possibility that people not only share information with others, but also they expose personal information unintentionally such as place to live, phone number, date of birth, and more. This will be serious problem if someone has impure mind. It is actually happening in cyber-stalking, offline stalking or others. There are also many spam messages in SNS because of the fact that information in SNS spread much faster than any other media and it is easy to send a message to others. In other words, SNS provides vast backbone environment to spammers to hunt normal pure users. In order to prevent information leakage and detect spam messages, many researchers traditionally have been studied for monitoring email systems, web blogs, and so on. In this paper, we dealt with text message data in Twitter which is one of the most popular social network services over the world in order to reveal various hidden patterns. Twitter data is severely dangerous to organizations and more is that anyone who has Twitter account can access to any users by "following" function. The following function does not require permission from the requested person to confirm to ready their timelines. This study will be focused on the user to whom exchange text messages and what types of information they reciprocated with others by monitoring 50 million tweets on November in 2009 which was collected by Stanford University.

Keywords: Information flow, Social network services, Information leakage, Twitter.

* Corresponding author.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 253–260, 2013.

1 Introduction

People are living in the place to find and share information with no constrains to time or location due to the huge enhancements of wireless internet infrastructure and Smartphone devices. We used to have to go back to home or internet cafe to search information or upload photos in very few years ago. However, we no longer have to go back to place to find desktop which has an internet access. We are simply able to obtain and share diverse information using Smartphone via mobile web browser or social network service (SNS) platform. Traditionally, the Web provides convenient and useful services to find information, knowledge, and more. People are willing to upload what they have been experienced during their travel or knowledge from their researches. Despite of this great convenience, there is a high possibility of personal private information leakage. The problem is that this personal information is leaking more seriously due to SNS. SNS is an online web platform to provide social activities among people. They share interests, activities, knowledge, events and more to strengthen their social relation with others in anytime and anywhere. There was a popular event make Twitter[2] got famous after U.S. Air ways jet crashes into the Hudson River on 15th of January 2009. The first photograph of this crash had appeared on Twitter earlier than any even local news media arrived at the accident place. This event brings an aspect that Twitter is not just a social web page but it is one of media. People got fascinated by this event so the popularity of Twitter was increased dramatically and it sinks into our life. The fastest social media is not always positive to people. Because of the fact that information on Twitter spread within a few second to all over the world, this might bring big obstacle to us all. The main reason why Twitter data is severely dangerous to organizations and more is that anyone who has Twitter account can access to any users by "following" function. The following function does not require permission from the requested person to confirm whether he/she will grant authority to read their timeline (or text message with others) or not. If user *A* send a following request to user *B*, user A automatically will be confirmed that he hereby can read timelines of user B based on the Twitter policy.

Let assume that user *C* sends a message to their friends to share information that tomorrow is his/her birthday. Or user *D* sends a message to celebrate his/her friend's birthday. In this case, although the date of birth is highly related to personal information, people normally expose precise date of birth unintentionally. The problem here is that as long as user *E* is following user *C* or *D*, user *E* is able to acquire personal text messages between user *C* and *D*. Moreover, people send a text to others with their real name or place to live. This is the main reason why we want to monitor personal information leakage on Twitter. Many researchers believe that SNS has great potential to reveal unknown personal attitude or sentiments but it still has an information leakage problem. This will be a serious problem if someone who has impure mind track personal information for cyber stalking. In order to prevent this problem on Twitter, we dealt with text message data in Twitter to reveal various hidden patterns related to personal information. This study will be focused on the user

[2] http://twitter.com

to whom reciprocated with by monitoring 50 million tweets on November in 2009 which was collected by Stanford University. We defined simple syntactic patterns to uncover date of birth in human written text messages.

The reminder of the paper is organized as follows: Section 2 describes related works; Section 3 explains a method for monitoring personal information leakage on Twitter based on syntactic patterns; Section 4 gives example for tracking breaking-events using Twitter; and finally Section 5 presents a conclusion to this work and makes suggestions for the future work.

2 Related Works

Digital personal information or personal identifiable information (PII) is always secured from other users. PII is information to uniquely identify or distinguish a single person identity. Full name, national identification number, driver license number, credit card number, date of birth, and more are commonly used to distinguish individual identity [1]. For example, if there is person who wants to withdraw huge amount of money from his/her bank accounts, he/she will be asked for presenting a valid PII to authenticate his/her identity. Moreover, when people forgot passwords for certain webpage, he/she will be asked for input PII data such as date of birth or email address. This digitized data can be easily duplicated to others and it tends to be exposed to others more readily than traditional physical resources [2]. This can be a serious crime if someone who has impure mind obtains digital PII intentionally or even accidently.

Over the years, many researchers have been studied for prevent personal information leakage not only in Internet web pages but also in SNS. Traditionally, packets which contain encrypted messages considered as an important factor to improve personal information security by monitoring transferring packets in networks [3]. Moreover, there was a research proposed a model to support a supply chain to make understand how confidential information of companies may be leaked using a conceptual model [4]. In order to infer private information in SNS, authors in [5] studied Facebook[3] data based on Naïve Bayes classification to predict privacy sensitive trait information among users. This research analyzed links among users to determine that personal information can be leaked to unknown person. [6] presented a new architecture for protecting personal private information published on Facebook for mitigating the privacy risks. Moreover, there was another research to trace social footprint of user's profile in order to uncover the fact that diverse personal information is leaking on multiple online social networks such as Flickr[4], LiveJournal[5], and MySpace[6] [7]. There is a big issue we have to give great attention that is personal information in SNS is leaking involuntarily [8]. People have been starting to take good care of preventing their personal information leakage when they

[3] http://facebook.com
[4] http://flickr.com
[5] http://livejournal.com
[6] http://myspace.com

make web documents. However, the problem is happened in SNS that SNS is full of freedom and enough metadata to infer someone's personal information. Users in SNS are likely to expose their private information unintentionally. This is why this paper focused on text messages in timeline on Twitter to reveal the fact that people reciprocate their personal information with others without any attention. Even they do beware of it, private information is leaking unintentionally.

3 Monitoring Personal Information Leakage on Social Network Services

This section describes a method for monitoring personal information leakage on Twitter. Twitter is a one of the most popular online SNS and microblogging service to share information by sending text-based messages restricted to 140 characters, known as "tweets" [9, 10]. Twitter users can freely follow others or are followed in contrary to most online SNS, such as Facebook or MySpace. Most of the SNS requires permission when users want to access others social web pages but not in Twitter. According to the Twitter policy, being a follower on Twitter means that users are able to obtain all the tweets from those users are following [9]. This issue guarantees a freedom of information sharing among anyone. However, personal information does not be exposed to everyone. In order to protect user personal information from unknown third parties, we define syntactic patterns to detect date of birth in human written text messages in Twitter. Let us assume that when people celebrate someone's birthday by a text message via Twitter or other SNS, the text message normally includes given keywords "birthday," "b-day." According to this assumption, we can obtain tweets which contain those keywords from huge amount of Twitter data set by simple text matching approach. The Twitter data set (8.27GB) which was collected by Stanford University [11] consists of time, user and tweet message information described in following Table 1.

Table 1. Examples of the Twitter data set

Type	Information
T	2009-11-02 14:49:31
U	*http://twitter.com/jhernandez48*
W	alright well off to cal class, I agree w/ Valentine on mike and mike but its a good things he is not the Philies manager, so oh well
T	2009-11-02 14:49:31
U	*http://twitter.com/kamanax*
W	This is the month of months and the week of weeks... Looking forward to the celebraaa on Friday, satday and Sunday! (cont) http://tl.gd/qkuu
T	2009-11-02 14:49:31
U	*http://twitter.com/koreainsider*
W	freezing waiting for bus, sweating on the bus and freezing again outside...I love that winter is here ?
…	…

T means the time when a given tweet was uploaded on Twitter and U indicates the Twitter user id who wrote the tweet. W represents the tweet at time T by user U.

In order to uncover the pattern hided in human natural language on Twitter, we simply extract tweets only include "birthday" and "b-day" from the data set. The size of extracted tweets was only approximately 30MB consist of around 189 thousand tweets. The following Table 2 shows examples of the extracted tweets.

Table 2. Examples of the extracted tweets from Twitter data set

Type	Information
T	2009-10-31 23:59:58
U	*http://twitter.com/nadiamaxentia*
W	@Anissarachma happy birthday to you, happy birthday to you, i wanna 'traktir' dont forget it, happy birthday to you. Haha
T	2009-11-01 00:01:45
U	*http://twitter.com/1crazy4justinb*
W	@justinbieber today is my b-day and it would mean the world to me if you told me happy birthday i love u!MAKE MY DREAM COME TRUE!<3
T	2009-11-01 00:10:04
U	*http://twitter.com/rgttos*
W	Kyny td udaah yaaa va, hahaha RT @virania: @rgttos happy birthday hhaaaa
…	…

Twitter has several functions such as 'RT,' '@,' '#,' and more. 'RT' means retweet, '@' indicates specific person (user id) whom user wants to send a tweet message and '#' is a hashtag that represents keywords or topics of tweets. We hereby define patterns to infer personal data of birth from the extracted tweets as described in Table 3. In order to define those patterns, we checked entire 189 thousand tweets by manually.

Table 3. Syntactic patterns for detecting date of birth

Index	Pattern information
1	***someone*** **(name or user ID) happy birthday (b-day) to** ***someone*** **(name, userID, or pronoun)**
ex)	@Anissarachma happy birthday to you
	Frank happy happy birthday from my heart
	happy birthday to my cousin Nathan and @Franklero
2	***someone*** **(name or userID) happy birthday (b-day)** ***someone*** **(name, userID, or pronoun)**
ex)	@LauraThomas34 happy birthday friend
	@kerronclement happy birthday
	Lolo happy birthday
3	***date*** **(such as today or tomorrow) is** ***possessive*** **birthday (b-day)** or **it is** ***possessive*** **birthday**
ex)	@justinbieber today is my b-day
	@JackAllTimeLow today is my birthday loooooh
	@Teairra_Monroe Its my birthday and
4	**wish** ***someone*** **(name, user ID, or objective) a birthday**
ex)	@krystyl: On the east coast - its @drew's birthday! Everyone wish him happy bday
	will you please wish me a happy birthday
	@MixMastaMario I wish her a happy birthday

According to patterns described in Table 3, pattern 1 and 2 is for celebrating someone's birthday but pattern 3 is for celebrating his/her own birthday. Pattern 4 is for celebrating not only for themselves but also others. In the case of pattern 1 and 2, the first "*someone*" indicates name of user, user ID, or @user ID. The second "*someone*" represents name of user, user ID, @user ID, or pronouns. In order to detect name of user in text messages, we collect every name lists of boys and girls from the website[7] which contains 10,532 names. Moreover, it is easy to detect user ID due to the fact that users on Twitter are most likely to add "@" function when they send a text message to others. Therefore, if tweet is satisfied with pattern 1 and 2 when user ID comes at first, the date when this tweet was uploaded will be the date for use ID's birthday. In case for pattern 3, the important factors to determine which date is the date of someone's birthday are *date* and *possessive* words. The *date* can be one of words such as today, tomorrow, or specific date and *possessive* represents words e.g. my, his, her, *name*'s, and *user ID*'s. "*Someone*" in pattern 4 can be a name, user ID, or objective word such as "me, her, him, etc." In order to detect personal date of birth on Twitter, we developed simple extraction program based on above patterns using Python.

4 Experiment

In order to protect user personal information such as date of birth on Twitter, we conducted simple experiment using syntactic patterns described in Table 3 based on following Fig 1. The test Twitter data set which only includes "birthday" and "b-day" words contains 189,247 tweets with time, user address, and text message.

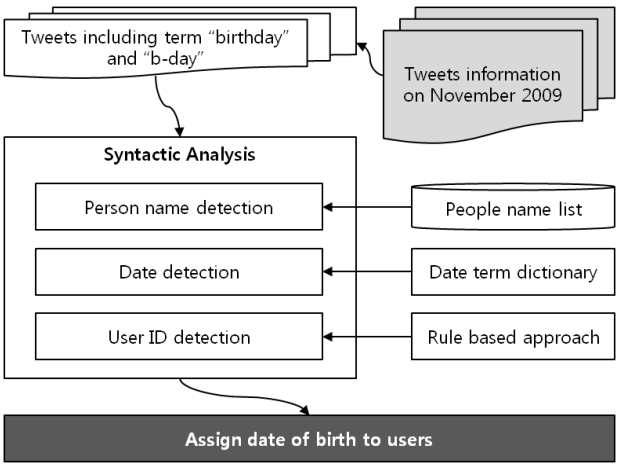


Fig. 1. Personal date of birth detection process

[7] http://www.momswhothink.com

Table 4. Example of experiment results

User	Date of birth	User	Date of birth
JanetRN	1st of November	davematthewsbnd	2nd of November
Larry	1st of November	sethredcast	2nd of November
1crazy4justinb	1st of November	Queen_MuLa_BaBy	2nd of November
dhilaloma	1st of November	ly_dalena	2nd of November
joseph	1st of November	dhardiker	2nd of November
Robert C Pernell	1st of November	glennmc	2nd of November
Brian Walker	1st of November	DJAYBUDDAH	2nd of November
rgttos	1st of November	TwentyFour	2nd of November
twephanie	1st of November	SoleHipHop	2nd of November
…	…	…	…

According to the proposed detection process in Fig 1, we can infer the user's date of birth as described in Table 4.

Let us assume that we have a tweet "RT @rhonda_ Happy birthday @JanetRN" by user apostlethatroks. This given text message indicates that user apostlethatroks send a celebration message which was originally written by user rhonda to user JaneRN. In other words, it was the birthday of user JanetRN. However, there is a problem that it is not guarantee the extracted date of birth is the precise date of user's birthday due to the fact that the proposed syntactic patterns cannot represent various human written text messages. In order to test how the proposed method can detect personal date of birth precisely, we randomly selected a hundred results to compare its accuracy manually. As a result, we can infer 61 percent of date of birthday from test data set with 75 percent of accuracy rate despite of fact that we only defined four kinds of syntactic patterns.

5 Conclusion and Future Works

In this paper, we proposed a method for monitoring personal information leakage on Twitter by inferring date of birth using proposed syntactic patterns. People can upload diverse information on social network services with no constrain to time or location. This fact brings great convenience to us all but it is still challenging issue. The serious problem is that people are likely to expose their personal information unintentionally via SNS. Therefore, we proposed simple syntactic patterns to give an idea to protect personal private information. Considering that only four kinds of patterns were applied, we believe that the inference rate from test data set is acceptable. If we can define syntactic patterns in detail, the results will be much better than this work. However, it is difficult to determine users when test tweets have pronouns, possessive, or objective words. In the nearest future, we are planning to apply named entity disambiguation approach to enhance the performance.

Acknowledgments. This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea (NRF) through the Human Resource Training Project for Regional Innovation.

References

1. Krishnamurthy, B.: I know what you will do next summer. ACM SIGCOMM Computer Communication Review 40(5), 65–70 (2010)
2. Yim, G., Hori, Y.: Guest Editorial: Information Leakage Prevention in Emerging Technologies. Journal of Internet Services and Information Security 2(3-4), 1–2 (2012)
3. Choi, D., Jin, S., Yoon, H.: A personal Information Leakage Prevention Method on the Internet. In: IEEE 10th International Symposium on Consumer Electronics, pp. 1–5 (2006)
4. Zhang, D.Y., Zeng, Y., Wang, L., Li, H., Geng, Y.: Modeling and evaluating information leakage caused by inference in supply chains. Computers in Industry 62(3), 351–363 (2011)
5. Lindamood, J., Heatherly, R., Kantarcioglu, M., Thuraisingham, B.: Inferring private information using social network data. In: Proceedings of the 18th International Conference on World Wide Web, pp. 1145–1146 (2009)
6. Lucas, M.M., Borisov, N.: FlyByNight: Mitigating the Privacy Risks of Social Networking. In: Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society, pp. 1–8 (2008)
7. Irani, D., webb, S., Pu, C., Li, K.: Modeling Unintended Personal-Information Leakage from Multiple Online Social Networks. IEEE Internet Computing 15(3), 13–19 (2011)
8. Lam, I.F., Chen, K.T., Chen, L.J.: Involuntary Information Leakage in Social Network Services. In: Proceedings of the 3rd International Workshop on Security: Advanced in Information and Computer Security, pp. 167–183 (2008)
9. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a Social Network or a News Media. In: 19th International Conference on World Wide Web, pp. 591–600 (2010)
10. Java, A., Song, X., Finin, T., Tseng, B.: Why We Twitter: Understanding Microblogging Usage and Communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56–65 (2007)
11. Yang, J., Leskovec, J.: Patterns of Temporal Variation in Online Media. In: ACM International Conference on Web Search and Data Mining, pp. 177–186 (2011)

On the Efficiency Modelling of Cryptographic Protocols by Means of the Quality of Protection Modelling Language (QoP-ML)

Bogdan Ksiezopolski[1,2], Damian Rusinek[2], and Adam Wierzbicki[1]

[1] Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
[2] Institute of Computer Science, Maria Curie-Sklodowska University,
pl. M. Curie-Sklodowskiej 5, 20-031 Lublin, Poland

Abstract. The problem of efficiency in the IT systems is now widely discussed. One of the factors affecting the performance of IT systems is implementation and maintaining a high level of security. In many cases the guaranteed security level is too high in relation to the real threats. The implementation and maintenance of this protection level is expensive in terms of both productivity and financial costs.

The paper presents the analysis of TLS Handshake protocol in terms of quality of protection performed by the Quality of Protection Modelling Language (QoP-ML). The analysis concerns efficiency.

1 Introduction

In the design of teleinformatic systems, the analyst must consider the system performance. One of the aspects which influence system performance is its security. System security is represented by means of the security attributes [7] which precisely define the system security requirements. Finally, the security requirements are guaranteed by using different types of security measures [5] which are realized by means of cryptographic protocols. Cryptographic protocols can be run with different parameters which affect system performance [6]. Security and efficiency analysts must decide which security measures and efficiency parameters should be used for the protocol realization and whether the selection is sufficient. Such an approach can be achieved by means of the Quality of Protection systems where the security measures and efficiency factors are evaluated according to their influence on the system security and performance.

1.1 Related Work

In the literature the security adaptable models are introduced as the Quality of Protection (QoP) models [4, 5, 8–10, 12, 13]. S.Lindskog and E.Jonsson attempted to extend the security layers in a few Quality of Service (QoS) architectures [9]. Unfortunately, the descriptions of the methods are limited to the

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 261–270, 2013.

confidentiality of the data and based on different configurations of the cryptographic modules. Y.Sun and A.Kumar [13] created QoP models based on the vulnerability analysis which is represented by the attack trees. The leaves of the trees are described by means of the special metrics of security. These metrics are used for describing individual characteristics of the attack. In the article [5] B.Ksiezopolski and Z.Kotulski introduced mechanisms for adaptable security which can be used for all security services. In this model the quality of protection depends on the risk level of the analysed processes. A.Lua et al [10] presents the quality of protection analysis for the IP multimedia systems (IMS). This approach presents the IMS performance evaluation using Queuing Networks and Stochastic Petri Nets. E. LeMay et al [8] created the adversary-driven, state-based system security evaluation, the method which quantitatively evaluates the strength of systems security. In the article [12] D.C. Petriu et al present the performance analysis of security aspects in the UML models. This approach takes as an input a UML model of a system designed by the UMLsec extension [2] of the UML modelling language. This UML model is annotated with the standard UML Profile for schedualability, performance and time and is then analysed for performance. In the article [4] B.Ksiezopolski introduced the Quality of Protection Modelling Language (QoP-ML) which provides the modelling language for making abstraction of cryptographic protocols that put emphasis on the details concerning quality of protection. The intended use of QoP-ML is to represent the series of steps which are described as a cryptographic protocol. The QoP-ML introduced the multilevel [15] protocol analysis that extends the possibility of describing the state of the cryptographic protocol.

In the QoP-ML the time of analysis is an important factor. The analysis engine is the part of the core system so the analysis can be performed in real time systems. In the article we present the new method and construction for modelling the efficiency of cryptographic protocols by means of the QoP-ML. The advantage of this approach is that the efficiency can be modelled simultaneously with security attributes and can be done for real time systems. For illustration of the QoP analysis process we choose one of the most popular cryptographic protocols - TLS [16]. In the article [4] the syntax, semantics and algorithms of the QoP-ML are presented.

2 Case Study: TLS Handshake Protocol

In this section we are going to present the case study of QoP modelling of TLS cryptographic protocol. We are analysing the two versions of the protocol, the first one with compression of the transmitted data and the second one without compression. The flow of the TLS Handshake protocol is realized in five steps and the scheme is presented in Fig. 1

Notation for Fig. 1:
PK_X - the public key of the X;
ID_{SX} - the id of the session X;
SK_X - the secret key of the X;
V_{TLS} - the version of TLS protocol;
$V_{TLS}(SET)$ - the established version of TLS protocol;
Cip_X - the available cipher suite for the session X;
Com_X - the compression method for the session X;
CA - the certificate authority;
$PK_X(cert) = (PK_X, ID_X, T)_{SK_{CA}}$ - the certificate of the X;
T - the timestamp;
ID_X - the id of the site X;
N_X - the nonce of the X.

1. $C \rightarrow S : ID_{S1}, V_{TLS}, Cip_1, Com_1, N_1$
2. $S \rightarrow C : V_{TLS}(SET), Cip_1(SET), Com_1(SET), N_2, PK_S(cert), Done(S)$
3. $C \rightarrow S : (K_1)_{PK_S(cert)}, ReadyEnc(C), Fin(C)$
4. $S \rightarrow C : ReadyEnc(S), Fin(S)$
5. $C \rightarrow S : (Data)_{K_1}$

Fig. 1. The protocol flow of the TLS Handshake protocol

The version of the TLS protocol presented in the Fig. 1 is the standard one. This protocol is fully analysed and described in [16].

The QoP analysis process includes the five steps: protocol modelling, security metrics definition, precess instantiation, QoP-ML processing and QoP evaluation. The following subsections describe these steps during modelling of the TLS protocol.

2.1 Protocol Modelling

In the first step one has to model all operations required in the TLS Handshake protocol. These operations are generally described in the protocol flow scheme (Fig.1). The complete QoP analysis of cryptographic protocols should contain many aspects like: the use of any security mechanism (not only cryptographic operation), key management operations, security policy management, legal compliance, implementation of the protocol and cryptographic algorithms, communication process, data storage and other factors which influence the system security. These aspects can be modelled by means of QoP-ML but this process is very complex and its presentation would have to be described on many pages. Therefore in the article we present one level analysis where only the efficiency factors which refers to cryptographic operation will be considered. The QoP analysis can refer to different security attributes and each of them must be proceeded according to the dedicated algorithms. In case of efficiency analysis we are focused on protocol time analysis which can be performed by means of availability algorithm which is introduced in [4].

The protocol modelling step includes the four operations [4]: function defining, equation defining, channels defining and protocol flow description.

Functions

For modelling of the TLS protocol we define the functions which refer to the cryptographic operations and affecting protocol efficiency. These functions are presented below. In the round bracket the description of these functions is presented.

```
fun id() (creating id of a session);
fun date() (create timestamp);
fun Vlist() (TLS versions list);
fun Clist() (creating ciphers list);
fun Comlist() (creating compression method list);
fun data() (prepare data);
fun set(X) (setup the X parameter);
fun info(X) (creating information message about X);
fun ReadyEncClient();
fun FinClient();
fun ReadyEncServer();
fun FinServer();
fun Done();
fun sk(id)[Av.:bitlength, algorithm] (compute secret key for id);
fun pk(sk)[Av.:bitlength, algorithm] (get public key from secret key);
fun cert(pk,id,t,ca)[Av.:bitlength, algorithm](compute certificate);
fun nonce() [Availability:bitlength, algorithm] (compute new nounce);
fun skey()[Availability:bitlength, algorithm] (compute symmetric key);
fun enc(data,key)[Availability:bitlength, algorithm, opt] (encrypt the data);
fun dec(data,key)[Availability:bitlength, algorithm, opt] (decrypt the data);
fun hmac(data)[Avail.: algorithm, block_size_in_MB] (hmac generation);
fun ver(X1,X2) (comparing X1 to X2);
fun com(data)[Avail.:data_type, blocksize_in_GB, algorithm](compression);
fun decom(data)[Av.:data_type, blocksize_in_GB, algorithm](decompression);
fun newstate(state) (state of the protocol);
fun st_active() (active state);
fun st_closed() (closed state);
```

Equations

After defining the functions one can describe the relations between them.

```
eq dec(enc(data,pk(SKid)),SKid) = data (asymmetric enc/dec)
eq dec(enc(data,K),K) = data (symmetric encryption/decryption)
eq ver(hmac(data),data) = true (verification of hmac digests)
eq decom(com(data)) = data (data compression/decompression)
```

Channels

In the presented example we define five synchronous channels.

```
channel ch1,ch2,ch3,ch4,ch5(100)[10 mbits];
```

Protocol Flow

The last and the most important operation during the modelling process is abstracting the protocol flow. In the presented case study we analyse two versions of the TLS protocol. In the Listing 1 the TLS client is modelled and in the Listing 2 the TLS server is modelled.

Listing 1. The client of TLS protocol modelled in the QoP-ML

```
host Client (rr)(*)
{
 #D1 = data();

 process C(ch1,ch2,ch3,ch4,ch5)
 {
  ID1 = id();
  V1 = Vlist();
  C1 = Clist();
  Com1 = Comlist();
  N1 = nonce()[256,Linux PRNG];
  M1 = (ID1, V1, C1, Com1, N1);
  out(ch1:M1);

  in(ch2:Y);
  PKScert=Y[4];
  K1=skey()[256,Linux PRNG];
  K1E=enc(K1,PKScert)[2048,RSA,pk];
  ReadyEC=info(ReadyEncClient());
  FinC=info(FinClient());
  M3=(K1E, ReadyEC,FinC);
  out(ch3:M3);

  in(ch4:Q);
  Status=newstate(st_active());

  subprocess Cv1(*)
  {
   D1Com=com(D1)[bin,1.14,Deflate];
   D1ComE=enc(D1Com,K1)[256,AES,CBC];
   D1all=hmac(D1ComE)[SHA1,1];
   M5=(D1ComE,D1all);
  }

  subprocess Cv2(*)
  {
   D1E=enc(D1,K1)[256,AES,CBC];
   D1all=hmac(D1E)[SHA1,1];
   M5=(D1E,D1all);
  }

  out(ch5:M5);
  Status=newstate(st_closed());
 }
}
```

Listing 2. The server of TLS protocol modelled in the QoP-ML

```
host Server (rr)(*)
{
 # S = id();
 # CA = id();
 # SKS=sk(S)[2048,RSA];
 # PKS=pk(SKS)[2048,RSA];
 # T1=date();
 # PKScert=cert(PKS,S,T1,CA)[2048,RSA];

 process S(ch1,ch2,ch3,ch4,ch5)
 {
   in(ch1:X);
   V1ok=set(X[1]);
   C1ok=set(X[2]);
   Com1ok=set(X[3]);
   N2=nonce()[256,Linux PRNG];
   DoneS=info(Done());
   M2=(V1ok,C1ok,Com1ok,N2,PKScert,DoneS);
   out(ch2:M2);

   in(ch3:Y);
   ReadyES=info(ReadyEncServer());
   FinS=info(FinServer());
   M4=(ReadyES, FinS);
   out(ch4:M4);

   Status=newstate(st_active());

   in(ch5:Z);

   subprocess Sv1(*)
   {
    K1E=Y[0];
    D1ComE=Z[0];
    D1all=Z[1];
    K1=dec(K1E,SKS)[2048,RSA,sk];
    D1ComEbis=hmac(D1ComE)[SHA1,1];
    Vres=ver(D1all,D1ComEbis);
    D1Com=dec(D1ComE,K1)[256,AES,CBC];
    D1=decom(D1Com)[bin,1.14,Deflate];
   }

   subprocess Sv2(*)
   {
    K1E=Y[0];
    D1E=Z[0];
    D1all=Z[1];
    K1=dec(K1E,SKS)[2048,RSA,sk];
    D1Ebis=hmac(D1E)[SHA1,1];
    Vres=ver(D1all,D1Ebis);
    D1=dec(D1E,K1)[256,AES,CBC];
   }

   Status=newstate(st_closed());
 }
}
```

To analyse them, one does not have to design these two versions separately, these two versions can be abstracted in one protocol flow. During defining the protocol instantiation, one can specify the parameters characteristic of specific versions of TLS Handshake protocol.

2.2 Security Metrics Definition

When modelling the protocol, the designer needs to define the security metrics for all functions connected with each security attribute which he wants to test. In the presented case study we test the availability of two different configurations of TLS Handshake protocol. Hence, we need metrics for all functions that may affect the availability. We have checked the execution times of operations used in the TLS protocol that may be configured (ie. compression, encryption).

Many of security metrics may be obtained from the benchmarks present in both official hardware specifications and literature [14]. However, some metrics may depend on the hardware on which protocol is executed [1]. Therefore, designers should be able to compute those metrics on hosts on which the protocol will be executed. In our case study we have applied commonly used software to compute metrics, so that everyone can compute them on their host.

For encrypting, decrypting (both symmetric and asymmetric) we have used the openssl program with speed library [11]. It executes the checked operation over and over for a period of time (ie. 10s) and returns the results which were converted to ms per byte.

In the case of compression and decompression we used *gzip/gunzip* to compute metrics. It contains the *zlib* library that has implementation of compression algorithm based on *deflate*. It is called the *standard reference implementation used in a huge amount of software*.

For the functions which generate the nonce and asymmetric keys we prepared the software which architecture is described in the article [4].

In order to compare the results from the QoP estimation with the real implementation, we have performed a test in which we copied an Linux MEPHIS iso file (1.14GB) using the scp program (secured copy). We configured the ssh connection (used by scp program) to use the same algorithms as we modelled in the presented case study. The ssh uses the TLS Handshake protocol for data transition.

In the QoP-ML the security metrics are defined by the operator `metrics` and the body of the metrics is closed in the curly brackets. Details about other operators used for defining security metrics can be found in [4]. The metrics for analysed versions of TLS protocol are presented on Listing 3.

Listing 3. The metrics for TLS protocol

```
metrics
{
 conf(host1)
```

```
{
 CPU = Intel Core i7-3930K 3.20GHz;
 CryptoLibrary = openssl 0.9.8o-5ubuntu1.2;
 OS = Ubuntu 11.04 64-bit;
}

data(host1)
{
 primhead[function][bitlength][algorithm][opt][Av:time(ms)];
 primitive[enc][2048][RSA][pk][0.049];
 primitive[dec][2048][RSA][sk][1.611];
 #
 primhead[function][bitlength][algorithm][opt][Av:time(mspB)];
 primitive[enc][256][AES][CBC][0.0000000049];
 primitive[dec][256][AES][CBC][0.0000000049];
 #
 primhead[function][algorithm][block_size_in_MB][Av:time(ms)];
 primitive[hmac][SHA1][1][2.475];
 #
 primhead[function][output_size:exact(B)];
 primitive[id][8];
 primitive[data][1224065679];
}

data+(host1.1)
{
 primhead[function][bitlength][algorithm][Av:time(ms)]\\
 [output_size:exact(B)];
 primitive[nonce][256][Linux PRNG][0.0025][8];
 primitive[skey][256][Linux PRNG][0.0025][8];
 #
 primhead[function][data_type][blocksize_in_GB][algorithm]\\
 [Av:time(ms)][output_size:ratio];
 primitive[com][bin][1.14][Deflate][31150][1:0.1];
 primitive[decom][bin][1.14][Deflate][6506][1:10];
}

set host Client(host1.1);
set host Server(host1.1);
}
```

2.3 Process Instantiation

During the process instantiation one can define the versions of the modelled protocol. In the presented example we set two versions of the TLS protocol (Listing 4), the first version with data compression and the second one without compression. In these versions two high hierarchy processes are executed: `host Client` and `host Server`.

In `version 1` inside the process `host Client`, the process `C` is executed (function - `run`) with the subprocess `Cv1`. Inside the process `host Server` the process `S` is executed with the subprocess `Sv1`. The TLS protocol versions can be modelled by defining the subprocess which will be executed in the specific protocol instantiation.

Listing 4. The process instantiation for TLS protocol

```
version 1
{
 run host Client(*)
 {
  run C(Cv1)
 }
 run host Server(*)
 {
  run S(Sv1)
 }
}
```

```
version 2
{
 run host Client(*)
 {
  run C(Cv2)
 }
 run host Server(*)
 {
  run S(Sv2)
 }
}
```

The second version of the TLS protocol is similar to the first one with one exception, the data is not compressed. The data processing without compression is modelled as the subprocesses `Cv2` and `Sv2` so the process `C` will execute the subprocess `Cv2` and the process `S` the subprocess `Sv2`.

2.4 QoP-ML Processing and QoP Evaluation

The final step in the QoP analysis process is QoP-ML processing and QoP evaluation which can investigate the influences of the security mechanisms for the system efficiency. The total execution time (T_{Total}) of the two analysed versions of the TLS protocol is calculated. For the first version of the protocol the $T_{Total} = 42.86\ s$. In the second, without data compression, the $T_{Total} = 5.79\ s$. The execution time for the second version of the protocol is 86.49% shorter than in the first version.

During the analysis one can notice that the first version of the TLS protocol is very inefficient in the case of transmitting the big binary file. The compression ratio for the binary file is only 10% and the time of compression has the largest contribution to the total execution time. The reason for using compression is to reduce the size of data thanks to which the execution time for the creating message authentication code and the transmission will decrease. Reducing the data is justified when the transmitted data is a text, then the compression ratio for the algorithm *Deflate* is about 70%.

3 The Runtime of TLS Protocol Implementation

For validation of the presented case study we compare the results of the protocol runtime estimated in the QoP-ML to runtime of an actual TLS protocol implementation. During the analysis we omit the computational overhead for packet transmission time, bit string comparison time, any hard drive operation time. The test was performed by means of the *scp* program which transmits the data using the TLS protocol. During the test we transmit the MEPHIS Linux iso file analysed in QoP-ML. The first and the second versions of the TLS protocol were executed 50 times and the average value was calculated with the standard deviation. In Tab. 1 we present the test results and the runtime estimated in QoP-ML. Comparing the results of the protocol runtime estimated in the QoP-ML to the runtime of the actual TLS protocol implementation, one can conclude that the protocol runtime estimated in the QoP-ML is in the range specified by the standard deviation. These results confirm the correctness of the efficiency modelling based on the QoP-ML approach.

Table 1. The TLS protocol runtime

	T_{Total} [s] QoP-ML estimation	T_{Total} [s] scp tests	standard deviation
Version 1	42.86	41.44	2.22
Version 2	5.79	6.47	0.80

4 Conclusions

The aim of this study was to present a new method and construction in new language QoP-ML [4] and show how to perform an efficiency analysis of cryptographic protocols. A full, multi-level cryptographic protocol analysis is very complex and exceeds the opportunity to be presented in this article. The performed study includes two selected versions of the TLS protocol and only cryptographic algorithms were taken into account. The QoP modelling language allows to analyse protocols in terms of different security attributes. This paper presents an analysis in terms of availability. Based on the algorithms presented in the article [4] we calculate the total protocol runtime for two versions of the TLS protocol. The protocol runtime estimated in the QoP-ML was validated by the real usage of the TLS protocol and confirms the correctness of modelling based on the QoP-ML approach.

The main feature of QoP-ML is that the cryptographic protocol can be analysed on different levels of security analysis. Owing to that, the QoP analysis can take into consideration many factors which influence the overall system security and efficiency.

Acknowledgements. Research partially supported by the grant "Reconcile: Robust Online Credibility Evaluation of Web Content" from Switzerland through the Swiss Contribution to he enlarged European Union.

References

1. Jaquith, A.: Security Metrics: Replacing Fear, Uncertainty, and Doubt. Addison Wesley (2007)
2. Jürjens, J.: Secure System Development with UML. Springer (2007)
3. Jürjens, J.: Tools for Secure Systems Development with UML. International Journal on Software Tools for Technology Transfer 9, 527–544 (2007)
4. Ksiezopolski, B.: QoP-ML: Quality of Protection modelling language for cryptographic protocols. Computers & Security 31(4), 569–596 (2012)
5. Ksiezopolski, B., Kotulski, Z.: Adaptable security mechanism for the dynamic environments. Computers & Security 26, 246–255 (2007)
6. Ksiezopolski, B., Kotulski, Z., Szalachowski, P.: Adaptive approach to network security. Communications in Computer and Information Science 158, 233–241 (2009)
7. Lambrinoudakis, C., Gritzalis, S., Dridi, F., Pernul, G.: Security requirements for e-government services: a methodological approach for developing a common PKI-based security policy 2003. Computers & Security 26, 1873–1883 (2003)
8. LeMay, E., Unkenholz, W., Parks, D.: Adversary-Driven State-Based System Security Evaluation. In: Workshop on Security Metrics - MetriSec (2010)
9. Lindskog, S.: Modeling and Tuning Security from a Quality of Service Perspective. PhD dissertation, Department of Computer Science and Engineering, Chalmers University of Technology, Goteborg, Sweden (2005)
10. Luo, A., Lin, C., Wang, K., Lei, L., Liu, C.: Quality of protection analysis and performance modeling in IP multimedia subsystem. Computers Communications 32, 1336–1345 (2009)
11. Openssl Project: `http://www.openssl.org/`
12. Petriu, D.C., Woodside, C.M., Petriu, D.B., Xu, J., Israr, T., Georg, G., France, R., Bieman, J.M., Houmb, S.H., Jürjens, J.: Performance Analysis of Security Aspects in UML Models. In: Sixth International Workshop on Software and Performance. ACM, Buenos Aires (2007)
13. Sun, Y., Kumar, A.: Quality od Protection(QoP): A quantitative methodology to grade security services. In: 28th Confrence on Distributed Computing Systems Workshop, pp. 394–399 (2008)
14. Szalachowski, P., Ksiezopolski, B., Kotulski, Z.: CMAC, CCM and GCM/GMAC: advanced modes of operation of symmetric block ciphers in the Wireless Sensor Networks. Information Processing Letters 110, 247–251 (2010)
15. Theoharidou, M., Kotzanikolaou, P., Gritzalis, S.: A multi-layer Criticality Assessment methodology based on interdependencies. Computers & Security 29, 643–658 (2010)
16. RFC 5246: The Transport Layer Security (TLS) Protocol v.1.2 (2008)

DiffSig: Resource Differentiation Based Malware Behavioral Concise Signature Generation

Huabiao Lu, Baokang Zhao, Xiaofeng Wang, and Jinshu Su

School of Computer, National University of Defense Technology, Changsha, China
{ccmaxluna,zhaobaokang}@gmail.com, {xf_wang,sjs}@nudt.edu.cn

Abstract. Malware obfuscation obscures malware into a different form that's functionally identical to the original one, and makes syntactic signature ineffective. Furthermore, malware samples are huge and growing at an exponential pace. Behavioral signature is an effective way to defeat obfuscation. However, state-of-the-art behavioral signature, behavior graph, is although very effective but unfortunately too complicated and not scalable to handle exponential growing malware samples; in addition, it is too slow to be used as real-time detectors. This paper proposes an anti-obfuscation and scalable behavioral signature generation system, DiffSig, which voids information-flow tracking which is the chief culprit for the complex and inefficiency of graph behavior, thus, losing some data dependencies, but describes handle dependencies more accurate than graph behavior by restrict the profile type of resource that each handle dependency can reference to. Our experiment results show that DiffSig is scalable and efficient, and can detect new malware samples effectively.

Keywords: Behavioral Signature, Anti-obfuscation, Scalable, Resource Differentiation, Iterative Sequence Alignment, Handle Dependency.

1 Introduction

Malware, short for malicious software, is software designed to disrupt computer operation, gather sensitive information, or gain unauthorized access to a computer system [1]. Malware includes computer viruses, worms, botnet, trojan horses, spyware, and so on. As we are know, Malware are the most serious threats in Internet for a long period. In fact, malware is the cause of most Internet problems such as spam e-mails and DoS (Denial of Service) [2].

The number of new malware samples is huge and growing at an exponential pace. Symantec observes 403 million new malware samples in 2011 [3], average 1.1 million malware samples per day. Efficiently analyzing such a scale malware samples and automated generating signatures for them is an interesting and challenging task.

Furthermore, obfuscated malware has become popular because of pure benefits brought by obfuscation: low cost and readily availability of obfuscation tools accompanied with good result of evading syntactic signature (e.g. byte-signature)

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 271–284, 2013.

based anti-virus detection as well as prevention of reverse engineer from understanding malwares' true nature. Obfuscated malwares use code obfuscation in software engineering to make the samples of a malware family look differently but have the same functionality.

The obfuscation techniques commonly used in malware obfuscation are Dead-Code Insertion, Register Reassignment, Instruction Substitution and Instruction Reordering [4]. Dead-code insertion is a simple technique that adds some ineffective instructions to a program. Register reassignment switches registers from generation to generation while keeping the program code and its behavior same. Instruction substitution evolves an original code by replacing some instructions with other equivalent ones. Instruction Reordering obfuscates an original code by changing the order of its independent instructions in a random way.

A system call (syscall) is how a program requests a service from an operating system's kernel, include hardware related services (e.g. accessing the hard disk, network card), creating and executing new processes, and so on [5]. The behavior of a program can be thought as its effect on the state of the system on which it executes. Thus, treating behavior in terms of the syscalls invoked by a program is reasonable, and this allows us to succinctly and precisely capture the intent of the malware [6]. syscalls are non-bypassable interfaces, so malware intending to unsafely alter the system will reveal its behavior through the syscalls that it invokes [7]. The current perfect way to describe malware behavior is the dependencies between syscalls [8].

The dependencies between syscalls can be presented by behavior graph, in which nodes are syscalls and an edge is introduced from a node x to node y when the syscall associated with y uses as argument some output that is produced by system call x. For example, if a program has a function merging two files (A,B) into a new file (C), the behavior graph for the function is shown in the Fig. 1. Ntreadfile(A, BufferA) uses the handle of file A as source, while handle of file A is produced by Ntopenfile(A), thus, there is an edge from the node Ntopenfile(A) to node Ntreadfile(A, BufferA). And Ntwritefile(C, BufferA +BufferB) writes a buffer which is computed from buffers produced by Ntreadfile(A, BufferA) and Ntreadfile(B, BufferB), and handle of file C produced by Ntcreatefile(C), therefore, there are a edge from nodes Ntreadfile(A, BufferA), Ntreadfile(B, BufferB) and Ntcreatefile(C) to node Ntwritefile(C, BufferA +BufferB) respectively. Dependent syscalls cannot be reordered without changing the semantics [8]. Thus behavior graph can defeat various obfuscated technologies naturally. Although effective, it is unfortunately too complex to be obtained and used, and it often requires cumbersome virtual machine technology [2].

A syscall sequence is the ordered sequence of syscalls that a process performs during its execution. Syscall sequence keep the orders assigned by behavior graph, that is saying, given an edge from a syscall x to syscall y in behavior graph, syscall x must be executed before syscall y. Besides the order of syscalls decided by behavior graph, we can gain handle dependencies between syscalls. In computer, a handle is an abstract reference to a resource, such as a file, a registry key, a process, a (network) socket, or a block of memory [9]. The syscall y is

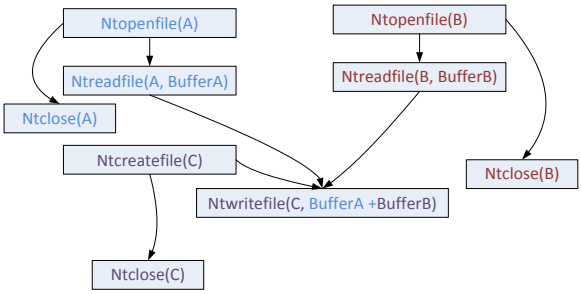


Fig. 1. Merging two files (A,B) into a new file (C)

dependent on syscall x on a handle only and only if the handle that is produced by syscall x and used by syscall y as input. we can figure out the dependency of syscalls on handles by simply checking the equality of handles in arguments of syscalls. In the Fig. 1, the dependency of Ntreadfile(A, BufferA)on Ntopenfile(A) is handle dependency. Therefore, the syscall sequence keeps the order and the dependency on handles presented by behavior graph, just losing other data dependencies that their data is modified between syscalls by mathematical or logical operations. The dependency Ntwritefile(C, BufferA +BufferB)on Ntreadfile(A, BufferA) and Ntreadfile(B, BufferB) are such data dependencies. All in all, syscall sequence is an approximate edition of behavior graph while evading strenuous information-flow tracking which is the chief culprit for the complex and inefficiency of graph behavior.

Dead-Code Insertion, Register Reassignment of malware obfuscation have no effect on syscall sequence. Instruction Substitution will evolves an original syscall sequece by replacing some syscalls with other equivalent syscalls. However, as limited syscalls in ordinary operating system, e.g., about 300 syscalls in Microsoft Windows OS, giving a syscalls block ,there is usually rare equivalent one to replace it. Therefore, syscall substitution has limited effect, we consider it in the future works. Instruction Reordering will lead to reorder the orders between independent syscalls. we present an Iterative Sequence Alignment (ISA) to perfectly defeat the syscalls reordering caused by malware obfuscation.

As executed in different environment or random tactics used, various samples of the same malware may bind the same handle to resource with different attributes, e.g. samples create files in different directories with the same purpose and writing to files with same contents, while looking the file path attributes of the handles, the handles are different and it is hard to relate them together. Thus, those differences hamper us to obtain handle dependency relation between different samples, and hamper us to gain a generalized behavior signature for those samples. We propose a resource differentiation scheme to seek common attributes of resource with same affect while hiding the differences of executing environment and the differences introduced by random tactics. The resource differentiation scheme aims to abstract the resources and to find handles with same usage from different samples. Thus it is called DiffHandle. DiffHandle is someway like DiffServ in IP QOS architecture of Internet.

After using DiffHandle to replace a handle with differentiated type of the resource that the handle reference to, and utilizing ISA to gain sub-signature (common non-consecutive ordered syscalls) set from syscall sequences of the same family, we obtain handle dependencies between syscalls in sub-signature set by backtracking those syscalls into original syscall sequences, at the same time, we gain the order restrictions between those syscalls. Finally, we generate a DiffHandle-signature consisting of syscalls that have handle dependencies –the handle can only relate to the designated type of resource– with others, and that have order restrictions between syscalls.

Comparing to behavior graph, DiffHandle-signature loses the data dependencies that their data is modified between syscalls by mathematical or logical operations, but, adds resource type (the type is defined by our resource differentiation scheme) restriction to handle dependencies. However, the most outstanding advantage of our method is avoiding complicated information-flow tacking, instead of it, gaining syscall sequence simply by hooking the SSDT (System Services Descriptor Table) table [11,16], even stirring advantage is that DiffHandle-signature can naturally be used as real-time detectors.

This paper proposes a resource differentiation based, anti-obfuscation, effective and efficient malware behavioral signature generation system, DiffSig, which is efficient, voiding complicated information-flow tacking which needs to use dynamic analysis infrastructure (virtual machine with shadow memory), generates a simplified but accurate malware behavioral signature for each malware family. In detail, the main contributions of this paper are as follows:

1. we propose a resource differentiation scheme (DiffHandle) to generalize syscall sequence by classifying resources while hiding the differences introduced by different executing environment or random tactics used by malwre.
2. we present an Iterative Sequence Alignment (ISA) to anti-obfuscation and gain sub-signature (common non-consecutive ordered syscalls) set from generalized syscall sequences.
3. we raise a concise but accurate behavioral signature presentation –DiffHandle-signature. Comparing to perfect but complicated behavior graph, DiffHandle-signature ignores some data dependencies that their data is modified between syscalls by mathematical or logical operations, but, adds resource type restriction to handle dependencies and avoids complicated information-flow tacking. It naturally suits to be used as real-time detector.

The rest of this paper is organized as follows. In section 2, we describe the system overview of DiffSig. And then we present the main three steps of DiffSig : DiffHandle, ISA and Backtracking refinement in sections 3, 4, 5 respectively. We validate our system using real malware samples in section 6. Finally, we conclude it and present future works in section 7.

2 System Overview

DiffSig consists of a kernel syscall monitor and three main steps to generate behavioral signature, as Fig. 2 shows.

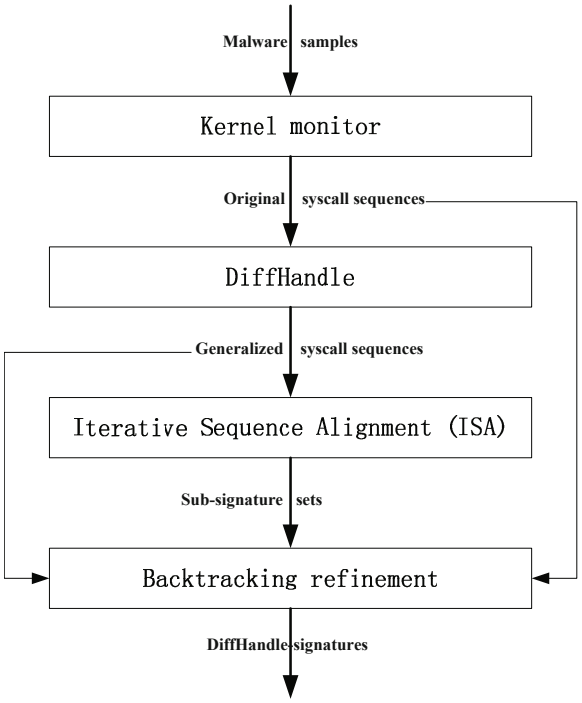


Fig. 2. Architecture of DiffSig

The kernel monitor collects syscall information for a designated process. In order to intercept and log system call information, the kernel monitor hooks the SSDT (System Services Descriptor Table) table [11,16] in Windows OS. It logs the timestamp, syscall name, arguments and return value of syscalls. Accordingly to timestamp of syscalls, all the syscalls of the process form a syscall sequence. This kind of sequence in which each syscall with raw arguments and return value is called original syscall sequence.

Resource Differentiation scheme (DiffHandle) classifies resource with same affect together, hiding the differences introduced by different executing environment or random tactics used by malware. For example, DiffHandle classifies file into: system files which generally locate under directory "C:\WINDOWS\system32"; private files which are operated only by a designated application, e.g. files under "C:\Program Files\iTunes" are private files of iTunes; and other types. The step of DiffHandle replaces a handle with differentiated type of the resource that the handle reference to, deletes the non-handle arguments of syscalls, and thus produces generalized syscall sequences in which each syscall is only with the types of resources it operates on.

Iterative Sequence Alignment (ISA) aims to extract all the common syscalls among various samples of the same family. We use sequence alignment to gain common syscalls. But, traditional sequence alignment can only gain partial

common syscalls because of syscall reordering, thus, we propose a new alignment method —Iterative Sequence Alignment (ISA). ISA applies multiple ways of traditional sequence alignment. The i^{th} way runs traditional sequence alignment on residual syscall sequences which consist of non-matched syscalls of $(i-1)^{th}$ way. ISA ends iterating when matched syscalls is less than predefined threshold. Some common non-consecutive ordered syscalls gained by each way are called as sub-signature. Multiple ways produce a sub-signature set.

Backtracking refinement finds out the handle dependencies that happens between syscalls in sub-signature set by putting those syscalls back to original syscall sequences of a same family. Backtracking refinement also obtains the order restriction among syscalls, especially, among syscalls from different sub-signatures. After the above three steps, we generate a concise but accurate DiffHandle-signature for each family.

DiffHandle-signature is formalized as a 3-tuple: $\{S, D, O\}$. $S = \{s_1, s_2, \cdots, s_n\}$ presents the common n syscalls extracted by DiffSig; $D = \{d_1, d_2, \cdots, d_m\}$ where $d_k = \{s_i, s_j, Type_{ij}\}(1 \leq k \leq m;\ s_i, s_j \in S)$ presents that there is a handle dependency which can only reference to resource type $Type_{ij}$ between syscall s_i and s_j; and the order restriction $O = \{o_1, o_2, \cdots, o_l\}$ where $o_k =< s_i, s_j > (1 \leq k \leq l;\ s_i, s_j \in S)$ presents that syscall s_i must be invoked before the syscall s_j.

3 Resource Differentiation Scheme (DiffHandle)

Resource Differentiation scheme (DiffHandle) classifies resource with same affect together, hiding the differences introduced by different executing environment or random tactics of malware. It is someway like DiffServ in IP QOS architecture in Internet. This step replaces a handle with differentiated type of the resource that the handle reference to, and delete the non-handle arguments of syscalls, produces generalized syscall sequences. We describe the differentiation scheme of three permanent resources: file, register key and network as follows.

3.1 File Differentiation Scheme

we classify all the files in Windows OS into four categories preliminarily according to their usage: System file, Private file, User file and Temporary file. Table 1 describes the four categories concisely. System files normally locate under directory "C:\WINDOWS\system32\", and are Windows system related services and data. Private files are normally only be operated by its owner application, e.g. files under the installation directory of a application is the private files of the application, thus, files under directory "C:\Program Files\" are private files to different applications, files under "C:\Documents and Settings\xxxx\Application Data" stores data of applications, and also are private files to corresponding application, another type of private file is the files under directory set by user for application , e.g. files under download directory of BitComet is private of BitComet. User files is files owns to a specified user or all users, such files

are created and modified during corresponding user active, files under directory "C:\Documents and Settings\Administrator\" are files of Administrator, and ones under "C:\Documents and Settings\All Users\" are shared among all users. Temporary files are intermediate files generated during applications execution, e.g. files under "C:\Documents and Settings\Administrator\Local Settings\Temp" are temporary files when user Administrator is active. Files not belong to above categories are treated as "other" file category.

Table 1. File Differentiation Scheme

TYPE	ID	Location	Description
System file	1	C:\WINDOWS\system32\	Windows System related execution and data
Private file	2	C:\Program Files\or others	Be operated only by its owner application normally
User file	3	C:\Documents and Settings\	Files owns to a specified user, such as Administrator
Temporary file	4	C:\Documents and Settings\ xxxx\Local Settings\Temp\	Intermediate files
Others	64	*	*

3.2 Register Key Differentiation Scheme

The most important type of register key is Autoruns. Autoruns are configured to run during system bootup or user login and they can make malwares run without any conscious or direct user interaction. Typical locations used by malwares are: "HKLM\System\Currentcontrolset\Services\%\Imagepath", "HKLM\Software\Microsoft\Windows\Currentversion\Run%", "HKLM\Software\Microsoft\Active Setup\Installed Components%", "HKLM\Software\Microsoft\Windows\Currentversion\Runonce%",and so on [15]. According to normal usage, locations are classified into seven categories preliminarily: Autoruns, HKLM\SYSTEM, HKLM\Software, HKLM\SECURITY, HKLM\SAM, HKEY_USER, HKEY_CLASSES_ROOT (HKCR). Table 2 gives a concise description to each of the seven categories. Register keys not belong to above categories are treated as "other" register key category.

3.3 Network Differentiation Scheme

We cluster network traffic into different categories according to higher-layer protocol (e.g. application-layer protocol). Network traffic are classified as Table 3. Application-layer types are identified by utilizing dynamic application-layer protocol identification technologies [14]. New category will be added as needed. The one not belong to above categories is treated as "other" network traffic.

All the resources can add new category as needed. We use longest prefix matching methods to identify which category the resource belongs to for File and Network traffic.

Table 2. Register key Differentiation Scheme

TYPE	ID	Location	Description
Autoruns	65	e.g. HKLM\Software\Microsoft\ Windows\CurrentVersion\Run	Shows what programs are configured to run during system bootup or user login
HKLM\SYSTEM	66	HKLM\SYSTEM \ HKEY_CURRENT_CONFIG\	information about the Windows system setup
HKLM\Software	67	HKLM\Software\	software and Windows settings, mostly modified by application and system installers
HKLM\SECURITY	68	HKLM\SECURITY\	linked to the Security database of the domain into which the current user is logged on
HKLM\SAM	69	HKLM\SAM\	reference all Security Accounts Manager (SAM) databases for all domains
HKEY_USER	70	HKEY_USERS\ HKEY_CURRENT_USER\	user profile actively loaded on the machine
HKCR	71	HKEY_CLASSES_ROOT	information about registered applications, e.g. file associations and OLE Object Class IDs
others	128	*	*

4 Iterative Sequence Alignment (ISA)

In this section, the most simple and basic sequence alignment, pairwise sequence, is introduced first; then we describe our Iterative Sequence Alignment (ISA).

4.1 Pairwise Sequence Alignment Algorithm

A pairwise sequence alignment is a matrix where one sequence is placed above the other to find and align common elements. Gaps ('-') are inserted to help in aligning matching characters. A mismatch occurs if elements in the same column are not identical. Fig. 3 shows results for pairwise sequence alignment between syscall sequences "O(A), R(A), CL(A),O(B), R(B), CL(B),CR(C), W(C), CL(C)" and "O(B), R(B), CL(B),O(A), R(A), CL(A),CR(C), W(C), CL(C)", where O(X) presents Ntopenfile(X), R(X) presents Ntreadfile(X, BufferX), CL(X) presents Ntclose(X), CR(X) presents Ntcreatefile(X), W(X) presents Ntwritefile(X, BufferX); assuming File A is a user file, File B is a temporary file, and File C also a user file, then the generalized syscall sequences is: "O(3), R(3), CL(3),O(4), R(4), CL(4),CR(3), W(3), CL(3)" and "O(4), R(4), CL(4),O(3), R(3), CL(3), CR(3), W(3), CL(3)". Since various samples of malware may bundling File A, B, C with different filename or file path, syscall generalization removes those differences and keep the rough profile of resource. In the rest of the paper, syscall sequences of the example are expressed in the same manner. The two syscall sequences in Fig. 3 are possible syscall sequences according to behavior graph described in Fig. 1. The common signature produced by the pairwise sequence alignment is "O(4)R(4)CL(4)*CR(3)W(3)CL(3)", which loses the operations on another user file.

Table 3. Network Differentiation Scheme

TYPE	ID	Description
DNS requests	129	*
HTTP-traffic	130	*
IRC-traffic	131	*
SMTP-traffic	132	*
P2P-traffic	133	*include some popular P2P types
FTP-traffic	134	include FTP-control, FTP-data and TFTP
SSL-traffic	135	associated to HTTPS connections
NetBIOS	136	*
TCP-traffic	137	TCP traffic other than those above-mentioned
UDP-traffic	138	UDP traffic other than those above-mentioned
ICMP-traffic	139	*
Listen	140	Listen on a port
others	192	*

4.2 Iterative Sequence Alignment (ISA)

In the above section, we introduce traditional one way sequence alignment which will lose some reordered syscalls. So, we propose a new sequence alignment mode, iterative sequence alignment, to defeat syscall reordering introduced by malware obfuscation. Fig. 4 shows results for iterative sequence alignment between syscall sequences "O(3), R(3), CL(3),O(4), R(4), CL(4),CR(3), W(3), CL(3)" and "O(4), R(4), CL(4),O(3), R(3), CL(3),CR(3), W(3), CL(3)". There are two ways alignment for the two sequences. The first way alignment is the same as traditional one way pairwise sequence alignment, and the second way alignment first obtain the mismatched elements in the previous way, then apply traditional sequence alignment on those mismatched elements . Thus, the sub-signature set is {O(4)R(4)CL(4)*CR(3)W(3)CL(3), O(3)R(3)CL(3)*}, it covers all the operations common in the two sequences. We define the regular iterative sequence alignment formally in the following.

O(3)	R(3)	CL(3)	O(4)	R(4)	CL(4)	-	-	-	CR(3)	W(3)	CL(3)
-	-	-	O(4)	R(4)	CL(4)	O(3)	R(3)	CL(3)	CR(3)	W(3)	CL(3)

Fig. 3. Example of Pairwise Sequence Alignment between Syscall Sequences

Definition 1. Iterative Sequence Alignment Problem
INPUT: A generalized sequence set $W = \{g_1, g_2, \cdots, g_{nw}\}$, a matched syscalls threshold to end the iteration: θ_{score}, and a noise threshold θ_{noise}.
OUTPUT: a sub-signature set Sig_w common among $(\lceil(1-\theta_{noise})(nw)\rceil)$ sequences of W. Sig_w satisfies the Formula 1 and $Score(SubSig_i)$ satisfies the matched syscalls threshold. $Score(SubSig_i)$ is the number of matched syscalls in i^{th} way sequence alignment.

$$\begin{aligned} &\textbf{Maximize} \quad Score(Sig_w) = \textstyle\sum_j (Score(SubSig_j)) \\ &\textbf{subject to} \quad Score(SubSig_j) \geq \theta_{score} \end{aligned} \tag{1}$$

The first way alignment:

O(3)	R(3)	CL(3)	O(4)	R(4)	CL(4)	-	-	-	CR(3)	W(3)	CL(3)
-	-	-	O(4)	R(4)	CL(4)	O(3)	R(3)	CL(3)	CR(3)	W(3)	CL(3)

The second way alignment:

	O(3)	R(3)	CL(3)	*
*	O(3)	R(3)	CL(3)	*

Fig. 4. Example of iterative Sequence Alignment between Syscall Sequences

5 Signature Refinement by Backtracking into Original Syscall Sequences

Result of iterative sequence alignment on the example is {O(4)R(4)CL(4)*CR(3) W(3)CL(3), O(3)R(3)CL(3)*}. However, we do not know whether the handle of CR(3) and the one of R(3) are same, syscall CR(3) and R(3) are handle dependable, and syscall R(3) must execute before W(3) from the result of ISA. This is what our signature refinement step do.

Definition 2. Signature Refinement Problem
INPUT: the selected $(\lceil(1-\theta_{noise})(nw)\rceil)$ generalized sequences $SG = \{sg_1, sg_2, \cdots, sg_{\lceil(1-\theta_{noise})nw\rceil}\}$, and corresponding $(\lceil(1-\theta_{noise})(nw)\rceil)$ original sequences $SO = \{so_1, so_2, \cdots, so_{\lceil(1-\theta_{noise})(nw)\rceil}\}$. Result of ISA: sub-signature set $SubSigSet = \{subsig_1, subsig_2, \cdots, subsig_{n\_isa}\}$ where each sub-signature $subsig_j = (syscall_{j1}, syscall_{j2}, \cdots, syscall_{jn_j})(1 \leq j \leq n\_isa)$ is some ordered syscalls that $syscall_{j(i-1)}$ must be invoked before $syscall_{ji}$.
OUTPUT: a DiffHandle-signature $FinalSig = \{FS, FD, FO\}$ that satisfies as Formula 2. DiffHandle-signature $FinalSig$ matches a original sequence sg_i means that there exists a map function which maps each handle type in FD to a handle

in sg_i and maps each syscall $s_x \in FS$ to a syscall s_y in sg_i, satisfying that for each element $\{s_i, s_j, Type_{ij}\} \in FD$, the mapper of s_i and the mapper of s_j are handle dependency in sequence sg_i and the type of the dependency is $Type_{ij}$, and for each element $< si, sj > \in FO$, the mapper of s_i stands before the mapper of s_j in sequence sg_i.

$$\begin{aligned} &\textbf{Maximize} \quad \|FD\| \\ &\textbf{subject to} \quad \forall syscall_k \in FS, \exists subsig_j \; that \; syscall_k \in subsig_j \\ &\textbf{and} \quad FinalSig \; matches \; all \; sg_i \; (1 \leq i \leq \lceil (1 - \theta_{noise})(nw) \rceil) \end{aligned} \tag{2}$$

6 Evaluation

In this section, we compare our DiffSig with previous famous system, Hamsa [12]. And verifies the effectiveness of DiffSig to detect new malware samples. we set following default parameters unless otherwise specified: matched threshold to end the iteration θ_{score} is 10, and a noise threshold θ_{noise} is 30%.

6.1 Data Set

We obtained a set of malware executables from mwanalysis.org [10] in the period from January 16, 2011 to March 21, 2011. According to the scan results of kaspersky anti-virus, we cluster the malware executables into different families. As a result, we gain 8 families and 331 malware executables described in Table 4 . The column "Family Name" is the scan results of kaspersky anti-virus pruning the variant name. Excepting detection result by heuristic methods, normally, the scan results consists of family name and variant name, while variant name is the string locating behind the last '.' in the scan result. E.g. in scan result of malware executable (MD5: 763ceb3a9127a0b9fac1bcf99a901d19): "Backdoor.Win32.Bifrose.usc", "Backdoor.Win32.Bifrose" is the family name, and "usc" is the variant name. The column "NUM of variants" presents the number of different variant names in a family. The column "NUM of executables" describes the number of different MD5 values in a family.

Since the kernel module hooking SSDT table not implemented yet, we utilize WUSStrace[13] to gain syscall sequences with detailed information. Besides running malware executables with WUSStrace to gain malicious syscall sequences, we also running normal application with WUSStrace to gain benign syscall sequences. Those normal programs contains firefox, iTunes, BitComet, Windows Office Word, realplay, FoxitReader, Skype, PartitionTableDoctor3.5, NokiaPCSuite7, PersonalBankPortal, Win32python2.7, and so on, totally 22 applications.

6.2 Performance of DiffSig

we compare our DiffSig with Hamsa [12] on above section mentioned malicious syscall sequences of malware executables and benign syscall sequences of normal

Table 4. The Families and Characteristics of Malware Executables

Family Name	NUM of variants	NUM of executables
Trojan-Downloader.Win32.CodecPack	13	48
Trojan-Dropper.Win32.VB	12	28
Trojan.Win32.VBKrypt	50	79
Worm.Win32.VBNA	14	75
Trojan.Win32.Refroso	21	27
Trojan.Win32.Cosmu	9	26
Trojan.Win32.Scar	7	21
Backdoor.Win32.Bifrose	19	27
Total	145	331

Table 5. The Performance of Our DiffSig Comparing to Hamsa

Family Name	Number of signatures		Detection rate		False Positive	
	DiffSig	Hamsa	DiffSig	Hamsa	DiffSig	Hamsa
Trojan-Downloader. Win32.CodecPack	1	3	95.8%	79.2%	0.045	0.135
Trojan-Dropper.Win32.VB	1	1	100%	78.6%	0	0
Trojan.Win32.VBKrypt	2	6	93.7%	88.6%	0	0
Worm.Win32.VBNA	2	5	93.3%	86.7%	0.045	0.18
Trojan.Win32.Refroso	1	4	92.6%	74.1%	0	0.090
Trojan.Win32.Cosmu	1	2	88.4%	76.9%	0	0
Trojan.Win32.Scar	1	—	90.4%	—	0	—
Backdoor.Win32.Bifrose	1	—	85.1%	—	0.045	—
average	1.25	3.5	92.4%	80.1%	0.0169	0.068

applications. Hamsa [12] extracts syscall token (consecutive syscalls) set,to form a signature as a set of common tokens, it aims to cover the most sequences under predefined false positive. Malicious syscall sequences are cut into two parts: one part as training set to generate behavioral signature and the other part as testing set to test the effectiveness of signature generated. Each part has half number of MD5 values and half number of variants.

Table 5 shows the performance of our DiffSig and Hamsa. The column "Number of signatures" describes the number of signatures generated by DiffSig and Hamsa respectively; The column "Detection rate" shows the detection ability of signatures to detect malware executables in testing set; The column "False Positive" presents the rate of treating normal application as malicious by

applying generated signatures to benign syscall sequences. From table 5, we can see that our DiffSig can generate signature for any family while Hamsa can not generate signature for Trojan.Win32.Scar and Backdoor.Win32.Bifrose because the two families have no invariable token in their syscall sequences. And we can see that our DiffSig produces fewer signatures since our Diffhandle-signature is more generalized, that the detection rate of our DiffSig is much higher and the false positive rate is much lower. In summary, our DiffSig generates more generalized and accurate signature, and our Diffhandle-signature is more suitable to anti-obfuscation than token set signature of Hamsa. Comparing the computing performance and detection performance with current behavior graph methods is our future work.

7 Conclusions

To surmount the complication of behavior graph but still to keep its effectiveness. We propose an anti-obfuscation and scalable behavioral signature generation system, DiffSig, which voids information-flow tracking which is the chief culprit for the complex and inefficiency of graph behavior, but describes handle dependencies more accurate than graph behavior by restrict the profile type of resource that each handle dependency can reference to. DiffSig is effective and efficient, generating concise but accurate signatures. Comparing with previous famous signature generation system, Hamsa, our DiffSig produces fewer signatures, detects more malware samples, and has lower false positive. Comparing the computing performance and detection performance with current behavior graph methods is our future works.

Acknowledgments. This work is supported by Program No. IRT 1012, the NSF of China Program No. 61202488, the Research Fund for the Doctoral Program of Higher Education of China No. 20124307120032, the NSF of China Program No. 61103194, and the NSF of China Program No. 61003303. We appreciate anonymous reviewers for their valuable suggestions and comments.

References

1. Wikipedia, `http://en.wikipedia.org/wiki/Malware`
2. Clemens, K., Paolo, M.C., Christopher, K., Engin, K., Xiaoyong, Z., Xiaofeng, W.: Effective and efficient malware detection at the end host. In: USENIX Security 2009, USENIX Press (2009)
3. Wikipedia, `http://www.symantec.com/threatreport/`
4. You, I., Yim, K.: Malware Obfuscation Techniques: A Brief Survey. In: 2010 International Conference on Broadband, Wireless Computing, Communication and Applications (2010)
5. Wikipedia, `http://en.wikipedia.org/wiki/System_call`
6. Fredrikson, M., Jha, S., Christodorescu, M., Sailer, R., Yan, X.: Synthesizing Near-Optimal Malware Specifications from Suspicious Behaviors. In: Proceedings of the 2010 IEEE Symposium on Security and Privacy (2010)

7. Srivastava, A., Lanzi, A., Giffin, J.: System Call API Obfuscation (Extended Abstract). In: Lippmann, R., Kirda, E., Trachtenberg, A. (eds.) RAID 2008. LNCS, vol. 5230, pp. 421–422. Springer, Heidelberg (2008)
8. Christodorescu, M., Jha, S., Kruegel, C.: Mining specifications of malicious behavior. In: Proc. of the 6th Joint Meeting of the European Software Engineering Conf. and the ACM SIGSOFT Symp. on The Foundations of Software Engineering (2007)
9. Wikipedia, http://en.wikipedia.org/wiki/Handle_(computing)
10. mwanalysis, http://mwanalysis.org/
11. Hoglund, G., Butler, J.: Rootkits: Subverting the Windows kernel. Addison Wesley Professional (2005)
12. Li, Z., Sanghi, M., Chen, Y., et al.: Hamsa: Fast Signature Generation for Zero-day Polymorphic Worms with Provable Attack Resilience. In: IEEE Symposium on Security and Privacy (2006)
13. http://code.google.com/p/wusstrace/
14. Dreger, H., Feldmann, A., Mai, M., Paxson, V., Sommer, R.: Dynamic Application-Layer Protocol Analysis for Network Intrusion Detection. In: 15th USENIX Security Symposium (2005)
15. Bayer, U., Habibi, I., Balzarotti, D.: A View on Current Malware Behaviors. In: 2nd USENIX Workshop on Large-Scale Exploits and Emergent Threats, LEET 2009 (2009)
16. Lanzi, A., Balzarotti, D., Kruegel, C., Christodorescu, M., Kirda, E.: AccessMiner: Using System-Centric Models for Malware Protection. In: CCS 2010. ACM Press (2010)

On Identifying Proper Security Mechanisms

Jakub Breier and Ladislav Hudec

Faculty of Informatics and Information Technologies,
Slovak University of Technology, Bratislava
{breier,lhudec}@fiit.stuba.sk

Abstract. Selection of proper security mechanisms that will protect the organization's assets against cyber threats is an important non-trivial problem. This paper introduces the approach based on statistical methods that will help to choose the proper controls with respect to actual security threats. First, we determine security mechanisms that support control objectives from ISO/IEC 27002 standard and assign them meaningful weights. Then we employ a factor analysis to reveal dependencies among control objectives. Then this knowledge can be reflected to security mechanisms, that inherit these dependencies from control objectives.

Keywords: Risk Evaluation, Information Security, Security Standards, Security Mechanisms, ISO/IEC 27002 standard.

1 Introduction

Operational cybersecurity is becoming more significant area of Computer Science. It is difficult to demonstrate a progress in this area, all the systems connected to the Internet are periodically under attack and the statistics about successful attacks still show the same ratio. In [4] authors analyze the progress in the automobile safety and compare it to the computer security. It is easier to eliminate known threats that do not change over time, as in the automobile industry the adversaries are natural laws that remain the same. In terms of computer security there are human adversaries that are evolving over time, therefore it is impossible to define static goals and to reach them.

We have to define security mechanisms that will help us to face the actual threats. There are numbers of these mechanisms, some are very effective, others have greater costs, but provide necessary industry protection controls and a number of them becoming useless as the Internet and computer networks evolve.

According to Baker et. al. [2], organizations tend to think more about quantity than quality. They are not aware, which mechanisms are the best for their purposes, so they often deploy as many as possible. Wrong mechanisms can actually add deficiencies to the system instead of increasing the security state.

In this paper we will try to find a way of defining proper security mechanisms for the organization. We will inspect the control objectives from ISO 27002:2005 [8] standard and assign one or more security mechanisms to each of 131 control objectives. We will also inspect dependencies between these mechanisms in order to correctly determine an evaluation criteria.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 285–294, 2013.

The main motivation for using security mechanisms is the clarity of their measurement. We cannot effectively measure the quality of control objectives implementation in the organization, but it is much easier with security mechanisms. If we have to use for example a mechanism called 'Implementation of authentication and authorization mechanisms - passwords, tokens, biometrics,' it is easy to decide whether it is fulfilled or not.

The overall goal is to determine whether the organization has satisfiable security controls in accordance to the ISO 27002:2005 standard and therefore it is able to demonstrate compliance with the ISO 27001:2005 [7] standard.

The rest of this paper is structured as follows. Section 2 provides an overview of a related work dealing with the problem of security mechanisms selection. Section 3 proposes our approach and describes methods used for choosing proper security mechanisms and for identification of relationships between them. Section 4 concludes this paper and provides a motivation for further work.

2 Related Work

In the field of security mechanisms and controls there are a few papers trying to propose an approach for their selection and implementation into the organization's information systems.

Singh ans Lilja [11] use Plackett & Burman (PB) design for determining the critical security controls [10]. This design requires minimum number of experiments to determine the effect of individual controls. For N controls it requires N+1 experiments. Each control can be implemented either as a low quality component or as a high quality component. These controls are then arranged in a matrix in a following way. Each row represents one experiment with numbers in columns either +1 or -1, indicating the control quality. Using these values together with the cost of each experiment we can determine the effect of particular security controls.

Authors compare 17 technical security controls, such as firewall, log analyzer, browser settings, etc.. They set up an experiment and provide an example of their method to prove its benefit in measuring impact of security enhancements.

Llanso [9] introduces CIAM - an approach that provides an initial prioritization of security controls. His approach uses data related to security incidents, vulnerabilities, business impact, and security control costs. He selects security controls from NIST 800-30 [12], assign them weights with support of security experts and estimate their efficiency against security breaches.

There are security standards, like the ISO 27002:2005 [8] or NIST 800-30 [12] that provide a database of security controls. But they fall short on choosing proper controls for the organization and on evaluation of quality of these controls. They also do not take in consideration relationships and dependencies over the security controls.

In [3] and [5] the authors deal with the similar problem looking at a more technical aspect - they introduce a scalable firewalling architecture based on dynamic and adaptive policy management facilities which enable the automatic

generation of new rules and policies to ensure a timely response in detecting unusual traffic activity as well as identify unknown potential attacks.

3 Methods

This section provides an overview of methods for selection of security mechanisms. Our approach emanates from the ISO/IEC 27002:2005 standard. We identify security mechanisms for each control objective from this standard and consider the importance and implementation quality of these mechanisms.

This section is divided into three subsections, the first one, the 3.1, provides the overview of the proposed weighting methods, the second one, the 3.2, describes relationships between security mechanisms and the last one, the 3.3, explains how to select and evaluate them.

3.1 Security Mechanism Weighting

Since there are many security mechanisms, an organization has to decide, which of them are useful and which are ineffective in contribution to its security goals.

There are eleven security clauses in the standard and each one is dealing with the different part of security, we have to use different types of security mechanisms. A NIST classification of security mechanisms constitutes three categories [12]. From our point of view, mechanisms used in our model also fits to one of these categories, therefore it is not necessary to use a new classification. Every security mechanism can be assign to one of the following groups: *Management, Operational or Technical.* It is much easier to measure the quality of the technical mechanisms, like firewall or intrusion prevention system, but it is impossible to quantify the quality of management or operational mechanisms, like information security policy. Because of character of ISO/IEC 27002 security clauses, that are mostly policy-based, we cannot measure all the mechanisms incorporated in the evaluation process automatically. But we can significantly improve the objectivity and simplicity of the evaluation.

We have to inspect them in two ways: how do they prevent against security breaches and how do they contribute to control objective fulfillment. Llanso [9] introduces an approach for selecting and prioritizing security controls (in the terminology of this paper, we use the term 'security mechanism' instead of the 'security control' because the latter term could indicate the usage of NIST 800-30 security controls). First, he computes weights of these controls, using three component weights - *Prevention, Detection and Response (P/D/R)* against an attack. The weight of a control i is computed by following equation:

$$RawWeighting_i = wP_i.owP_i + wD_i.owD_i + wR_i.owR_i \tag{1}$$

where overall weightings have values $owP_i = 0.5, owD_i = 0.25, owR_i = 0.25$, because prevention is more valuable than the other two. Control's contribution to these three actions (wP_i, wD_i, wR_i) are scores. These are determined by subject matter experts (SMEs) and each of them holds a value in interval $< 0, 1 >$.

After this step, he computes relative weighting as a ratio between one security control and all the other controls:

$$RelativeWeighting_i = \frac{RawWeighting_i}{\sum_{j=1}^{n} RawWeighting_j} \tag{2}$$

Then he is able to compute the priority, using relative weightings, scores, attack step frequencies, CVSS impacts and costs.

Since we do not have the cost dimension in our model, we will not use the whole prioritization approach. We will adopt the relative weighting process and adjust it in a meaning of contribution of security mechanisms to control objectives. We are not weighting these mechanisms with respect to possible attacks, but we are looking at how well do they assure the control objective function. So instead of *P/D/R* components we will use *Implementation, Maintenance and Policy (I/M/P)* components. The equation remains the same, just with another components and with another overall weightings:

$$RawWeighting_i = wI_i.owI_i + wM_i.owM_i + wP_i.owP_i \tag{3}$$

where overall weightings have values $owI_i = 0.6, owM_i = 0.20, owP_i = 0.20$. The implementation is the most important component, whithout them the maintenance components does not have a meaning, so we have to take them into consideration. That is why it has the highest value. The maintenance ensures the correct function of the control objective and the policy component specifies, whether the security mechanism supports also a formal policy. The relative weighting formula remains the same as in Equation 2.

Table 1 presents the "Controls against malicious code" control objective from "Communications and operations management" security clause. We assigned five security mechanisms to this control objective and used the same approach for determining weights as Llanso [9] did. We constituted a group of security professionals for this purpose, so they could discuss if the security mechanisms are properly assigned and what values can they achieve in each of three components. The last column represents a relative weighting of particular security mechanisms. Each component weight has a value on a discrete scale from 1 to 10, 1 means minimal importance, 10 is the most significant importance.

3.2 Correlation of Security Mechanisms

There is another dimension in the security mechanisms selection problem - a correlation between individual mechanisms. We cannot look on particular mechanisms as on the independent attributes, each one can affect the implementation of another one. It is usually better to have implemented for example three of them at an average implementation quality level than just one individual mechanism at a comprehensive level [6]. The cost is also a dimension that plays significant role in the above statement - the maximal quality of the implementation demands usually excessive resources. Commonly, it is more efficient to choose the way of implementing reasonable amount of mechanisms at a reasonable quality.

Table 1. Controls against malicious code

Security Mechanism	wI	wM	wP	RW
Implementing operating system policies prohibiting the use of unauthorized software, downloading unsigned executable files and working with other than data files on workstations without privileges.	9	5	7	**0.244**
Implementing strong account policies with separated privileges and clear accountability and non-repudiability.	7	3	9	**0.206**
Deployment of antivirus software on each system with the real-time check of unwanted code and periodical update of this software.	9	9	2	**0.238**
Ensuring that installed programs are up to date.	3	9	7	**0.156**
Providing business continuity plan - backuping and version management.	3	7	9	**0.156**

Since there are 131 control objectives and around 3-5 security mechanisms assigned to each of them at average, it would take a huge amount of time to determine correlation among each pair. We decided to choose a higher level of abstraction and to inspect a correlation between control objectives. We integrate this part of the model with the protection against security breaches, stated in the beginning of this section.

Verizon publishes Data Breach Investigations Reports [1] every year. It contains statistical records of incidents collected from various companies, divided into categories, providing detailed information about the overall state of the cyber security in our society.

We will use the Top 10 threat action types by number of breaches from this record and inspect, how particular control objectives provide prevention against these breaches. Ideal for this purpose is the Factor Analysis (FA) method, which describes variability among observed correlated variables. In this method, the measured variables depend on a smaller number of latent factors. Each factor can affect several variables in common, so they are known as *common factors*. Particular variables can be then represented as a linear combination of the common factors. The coefficients in this combination are known as *loadings*. FA can be used to reduce the redundant information contained in several correlated variables. However we will use it to reveal these correlations and to insert these dependencies in our measurement model.

To save the space, we will not use the whole set of control objectives, but we will pick one sample objective from each security clause. These are listed in Table 2 among columns in the following order: Information security policy document (CO_1), Confidentiality agreements (CO_2), Inventory of assets (CO_3), Information security awareness, education, and training (CO_4), Physical entry controls (CO_5), Disposal of media (CO_6), User password management (CO_7), Input data validation (CO_8), Reporting information security events (CO_9), Business continuity and risk assessment (CO_{10}), Protection of organizational records (CO_{11}). The evaluation is based on a discrete scale from 1 to 10, 1 means no protection

Table 2. Control objectives' protection against Top 10 security threats

Sec. Clause / Breach	CO_1	CO_2	CO_3	CO_4	CO_5	CO_6	CO_7	CO_8	CO_9	CO_{10}	CO_{11}
Keylogger/Form-grabber/Spyware	7	1	1	7	3	1	5	5	5	1	3
Exploitation of default or guessable credentials	7	3	1	8	3	1	9	1	4	1	3
Use of stolen login credentials	3	1	1	5	7	3	7	1	5	1	5
Send data to external site/entity	5	1	1	7	3	3	5	1	3	1	5
Brute force and dictionary attacks	7	1	3	9	5	3	9	1	5	1	5
Backdoor	5	3	1	7	5	1	5	5	5	1	3
Exploitation of backdoor or command and control channel	5	1	1	5	3	1	5	3	5	1	7
Disable or interfere with security controls	7	3	1	7	8	1	5	2	5	1	5
Tampering	8	3	1	8	3	1	1	1	5	1	3
Exploitation of insufficient authentication	7	3	1	8	7	1	5	1	3	1	5

and 10 means maximal protection. We can see that there are control objectives which are important in the view of these breaches, like Information security policy document, Information security awareness, education, and training, or User password management. On the other hand, there are objectives that have negligable importance, like Inventory of assets or Business continuity and risk assessment. The purpose of this evaluation is not to determine the control objectives' significance, but to reveal possible hidden relationships between them. Then we can reflect these findings in the security evaluation.

Now we can use the factor analysis on the matrix obtained from Table 2. Besides other important characteristics we get the Pearson's correlation matrix. In this matrix we can see dependencies between each two control objectives:

$$
\begin{array}{c}
 \\
CO_1 \\
CO_2 \\
CO_3 \\
CO_4 \\
CO_5 \\
CO_6 \\
CO_7 \\
CO_8 \\
CO_9 \\
CO_{10} \\
CO_{11}
\end{array}
\begin{array}{c}
\begin{array}{ccccccccccc}
CO_1 & CO_2 & CO_3 & CO_4 & CO_5 & CO_6 & CO_7 & CO_8 & CO_9 & CO_{10} & CO_{11}
\end{array} \\
\left(\begin{array}{ccccccccccc}
1 & 0.484 & 0.208 & 0.788 & -0.171 & -0.498 & -0.208 & -0.092 & -0.043 & -0.715 & -0.400 \\
0.484 & 1 & -0.333 & 0.410 & 0.263 & -0.655 & -0.273 & -0.063 & -0.124 & -0.333 & -0.469 \\
0.208 & -0.333 & 1 & 0.519 & 0.053 & 0.509 & 0.515 & -0.232 & 0.207 & -0.111 & 0.156 \\
0.788 & 0.410 & 0.519 & 1 & -0.073 & -0.054 & 0.127 & -0.265 & -0.254 & -0.573 & -0.473 \\
-0.171 & 0.263 & 0.053 & -0.073 & 1 & 0.103 & 0.139 & -0.190 & 0.033 & 0.404 & 0.255 \\
-0.498 & -0.655 & 0.509 & -0.054 & 0.103 & 1 & 0.417 & -0.456 & -0.135 & 0.509 & 0.307 \\
-0.208 & -0.273 & 0.515 & 0.127 & 0.139 & 0.417 & 1 & -0.190 & -0.056 & 0.212 & 0.128 \\
-0.092 & -0.063 & -0.232 & -0.265 & -0.190 & -0.456 & -0.190 & 1 & 0.432 & -0.232 & -0.267 \\
-0.043 & -0.124 & 0.207 & -0.254 & 0.033 & -0.135 & -0.056 & 0.432 & 1 & 0.207 & -0.097 \\
-0.715 & -0.333 & -0.111 & -0.573 & 0.404 & 0.509 & 0.212 & -0.232 & 0.207 & 1 & 0.156 \\
-0.400 & -0.469 & 0.156 & -0.473 & 0.255 & 0.307 & 0.128 & -0.267 & -0.097 & 0.156 & 1
\end{array}\right)
\end{array}
$$

Table 3 shows us the unrotated component matrix, consisting of three main factors. This matrix represents the significance of elements within each factor.

Table 3. Unrotated component matrix

	F_1	F_2	F_3
CO_1	0.858	0.313	0.048
CO_2	0.690	-0.145	-0.434
CO_3	-0.128	0.851	0.436
CO_4	0.693	0.720	-0.023
CO_5	-0.195	0.040	-0.303
CO_6	-0.727	0.540	-0.027
CO_7	-0.317	0.432	0.082
CO_8	0.176	-0.573	0.671
CO_9	-0.081	-0.188	0.413
CO_{10}	-0.720	-0.121	-0.218
CO_{11}	-0.506	0.059	-0.073

For better visualisation, the results are stated in Figure 1. By inspecting factor 1, we can see that it depends on the following control objectives: Information security policy document, Confidentiality agreements, Information security awareness, education, and training. It means that these control objectives are somehow bounded together from the view of security breaches. Factor 2 has higher loadings for control objectives Inventory of assets, Information security awareness, education, and training, and Disposal of media. The last factor has higher loading only for Input data validation, so we can say there will be no dependence emanating from this factor.

Obviously, the dependence cannot be determined only by mathematical methods because of the character of particular control objectives. For example, if we have an objective that supports implementation of antivirus software and the other objective, implementing periodical software updates, these are clearly highly correlated. However, we can say that the second one supports the first one highly, but it does not work in the opposite way. Software updates are not affected by implementation of antivirus software, so there is only one-way dependence. We have to use a group of security professionals to determine the character of dependencies.

We can explicate the results in the following way. Information security policy document is clearly an important control objective, Confidentiality agreements and Information security awareness, education, and training objectives depend on it. The latter two do not contribute to the first one, so there will be only one way correlation. Similarly, they do not have cross-dependency. Disposal of media and Inventory of assets are dependent on Information security awareness, education, and training. Disposal of media is also dependent on Inventory of assets. So we have five relationships in total, each of them is only one way dependence. Now we can use the correlation values to affect the evaluation of security mechanisms.

In Equation 4 we can see the evaluation of control objective i. S_{CO_i} is the score of control objective i, obtained by the evaluation, RW_{CO_i} is its weight, S_{CO_j} is the score of control objective j, that is correlated with i and COR_{ij} is

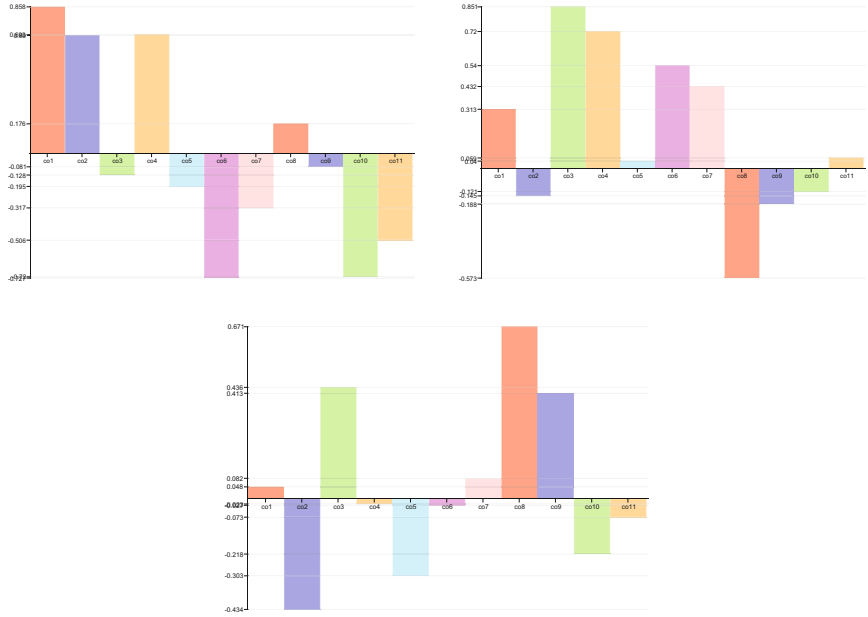


Fig. 1. Components and control objectives

the correlation between them. It is easy to see that the fraction can gain values from interval $< 0, 0.5 >$ and can significantly improve the value of the final score, if both the correlation and the correlated control objective's score are high.

$$FinalScore_{CO_i} = RW_{CO_i} * \left(S_{CO_i} + \frac{S_{CO_j} * COR_{ij}}{1 + COR_{ij}} \right) \tag{4}$$

The score of S_{CO_i} depends on evaluation of security mechanisms associated to the control objective i and it is the product of the security mechanism's weighting and its score. The calculation of S_{CO_i} is stated in Equation 5. The score of the security mechanism's implementation $(S_(M_j))$ is determined by security analyst and can have a value in interval $< 0, 1 >$, 0 means no implementation and 1 means that it is implemented well, tested and verified in a real environment.

$$S_{CO_i} = \sum_{j}^{n} S_{M_j} * RW_{M_j} \tag{5}$$

The evaluation of control objective's weight (RW_{CO_i}) is not in the scope of this paper, since we do not propose the complete evaluation model, we only designate a selection method for security mechanisms and identify their relationships and dependencies.

3.3 Security Mechanism - Selection Process and Evaluation

To summarize the previous sections, the assignment of security mechanisms to control objectives consists of following steps:

1. Assignment of the security mechanisms with respect to control objective's description. Usually there are three to five mechanisms supporting one control objective.
2. Definition of the weight of each mechanism - this parameter shows us, how important is this mechanism for the contol objective fulfillment. We use the I/M/P model with three weight components for this purpose. The sum of relative weights of mechanisms is 1.
3. Determination of weights of control objectives - these weights have to reflect the organization's security goals, for example necessary requirements for confidentiality, integrity or availability of organization's assets.
4. Estimation of dependencies with the factor analysis method, adjusted by the security analyst's judgement.

4 Conclusions

In this paper we proposed a way of choosing proper security mechanisms that will protect the organization's assets. We defined a method for the determination of importance of these mechanisms by assigning weights. These weights express, how well the particular mechanism contributes to the implementation, maintenance or policy fulfillment of the control objective, to which it was allocated. The whole model consists of about four hundred security mechanisms, that were allocated for each of 131 control objectives from the ISO/IEC 27002:2005 standard.

We also proposed the approach for determining relationships between security mechanisms. For this purpose we choosed the factor analysis, a statistical method that can reveal hidden correlation among observed variables. We explore these correlations on the control objectives layer, because inspecting every security mechanism would be exceedingly comprehensive and space consuming. The factor analysis gave us meaningful results that need to be further adjusted by security professionals to express the dependencies correctly.

Verendel [13] claims that quantitative security evaluation is still very unclear and it is almost impossible to validate the methods against empirical data. We would like to confute this claim by building a quantitative evaluation method that will be based on a number of smaller components, that work together and could be verified standalone. In this paper we presented the component dealing with the security mechanisms problem.

In the future, we would like to use the results of this work to construct a security evaluation system, that will measure the real security state in an organization by evaluating the quality of implemented security mechanisms.

Acknowledgment. The paper was prepared with partial support of research grant VEGA 1/0722/12 entitled "Security in distributed computer systems and mobile computer networks".

References

1. Baker, W., Hutton, A., Hylender, D., Pamula, J., Porter, C., Spitler, M.: 2012 Data Breach Investigations Report. Technical report, Verizon (2012)
2. Baker, W., Wallace, L.: Is information security under control?: Investigating quality in information security management. IEEE Security and Privacy 5(1), 36–44 (2007)
3. Castiglione, A., De Santis, A., Fiore, U., Palmieri, F.: An enhanced firewall scheme for dynamic and adaptive containment of emerging security threats. In: 2010 International Conference on Broadband, Wireless Computing, Communication and Applications (BWCCA), pp. 475–481 (November 2010)
4. Cybenko, G., Landwehr, C.E.: Security analytics and measurements. IEEE Security & Privacy 10, 5–8 (2012)
5. De Santis, A., Castiglione, A., Fiore, U., Palmieri, F.: An intelligent security architecture for distributed firewalling environments. Journal of Ambient Intelligence and Humanized Computing, 1–12 (2011)
6. Gordon, L.A., Loeb, M.P.: The economics of information security investment. ACM Trans. Inf. Syst. Secur. 5(4), 438–457 (2002)
7. ISO. ISO/IEC Std. ISO 27001:2005, Information Technology - Security Techniques - Information security management systems - Requirements. ISO (2005)
8. ISO. ISO/IEC Std. ISO 27002:2005, Information Technology - Security Techniques - Code of Practice for Information Security Management. ISO (2005)
9. Llanso, T.: CIAM: A data-driven approach for selecting and prioritizing security controls. In: 2012 IEEE International Systems Conference (SysCon), pp. 1–8 (March 2012)
10. Plackett, R.L., Burman, J.P.: The design of optimum multifactorial experiments. Biometrika 33(4), 305–325 (1946)
11. Singh, A., Lilja, D.: Improving risk assessment methodology: a statistical design of experiments approach. In: Proceedings of the 2nd International Conference on Security of Information and Network (SIN 2009), pp. 21–29. ACM, New York (2009)
12. Stoneburner, G., Goguen, A., Feringa, A.: NIST Special Publication 800-30: Risk Management Guide for Information Technology Systems. In: NIST (2002)
13. Verendel, V.: Quantified security is a weak hypothesis: a critical survey of results and assumptions. In: Proceedings of the 2009 Workshop on New Security Paradigms Workshop (NSPW 2009), pp. 37–50. ACM, New York (2009)

A Recovery Approach for SQLite History Recorders from YAFFS2

Beibei Wu, Ming Xu, Haiping Zhang, Jian Xu, Yizhi Ren, and Ning Zheng

College of Computer, Hangzhou Dianzi University, Hangzhou 310018
Jhw_1314@126.com,{mxu,zhanghp}@hdu.edu.cn

Abstract. Nowadays, forensic on flash memories has drawn much attention. In this paper, a recovery method for SQLite database history records (I.e. updated and deleted records) form YAFFS2 is proposed. Based on the out-of-place-write strategies in NAND flash memory required by YAFFS2, the SQLite history recorders can be recovered and ordered into timeline by their timestamps. The experiment results show that the proposed method can recover the updated or deleted records correctly. Our method can help investigators to find the significant information about user actions in Android smart phones by these history recorders, although they seem to have been disappeared or deleted.

Keywords: Digital forensics, Android, YAFFS2, SQLite, Recovery.

1 Introduction

With the growth of Android smart phones, the need for digital forensics in this area has shown a significant increase. For the small size, and fast running speed, SQLite is widely used in application software that needs to save simple data in a systematic manner or adopted into embedded device software. In the Android, a large amount of user data is stored in the SQLite database, such as short messages, call logs, and contacts [1]. Considering that deletion of data is frequently practiced in order to manage storage space or to update with the latest data, acquiring deleted data information is equally as important as retrieving undamaged information from the database.

Although since the beginning of 2011 with version Gingerbread (Android 2.3), the platform switched to the EXT4 file system, there are still many devices in use running a lower version than 2.3 and using YAFFS2. Therefore, insights into the amount and quality of evidence left on YAFFS2 devices are still of major interest. Compared to databases in some other physical environments, because of the write-once limitation of NAND flash and YAFFS2 page allocation mechanism [2-3], SQLite in restoring means are different. This paper proposed a recovery method of updated or deleted record for SQLite database based YAFFS2. Consequently, with sequence number and timestamps for SQLite database file, we can construct timeline of SQLite operating log and then analyze user actions.

2 Related Work

Researches of database record recovery are already begun in 1983 by Haerder [4]. He suggested that the deleted record can be recovered using transaction file. This method

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 295–299, 2013.
IFIP International Federation for Information Processing 2013

can be applied to traditional database on PC when the information of deleted record is included in transaction file.

A study conducted by Pereira [5] attempted recovering deleted records in Mozilla Firefox 3 using rollback journal file. In the paper, an algorithm to recover deleted SQLite entries based on known internal record structures was proposed and an exception was used that the rollback journal is not deleted at the end of each transaction when the database is used in "exclusive locking" mode.

While the objective of deleted data recovery in SQLite is on the line of Pereira's work, a tool using recovery method that approaches actual data files instead of a journal file that would offer improved practical availability was suggests by Sangjun Jeon [6]. They analyzed the file structure of SQLite database and proposed a method to recover deleted records from the unallocated area in page. However, this method is hard to recover deleted records because remained data are partial in deleted area and the length of each field is difficult to estimate. For the android phone that implemented wear-leveling by YAFFS2, the deleted records can be recovered from previous versions of file.

3 Recovery Elements of SQLite Deleted Records

From the "out-of-place-write" strategy of YAFFS2 and atomic commit in SQLite [7], a deleted SQLite file could be recovered. Therefore, the deleted records can also be restored. In YAFFS2, obsolete chunks can only be turned into free chunks by the process of garbage collection. Whenever one or more obsolete chunks exist within a block, the corresponding data is still recoverable until the respective block gets garbage collected. And from the perspective of the storage mechanism of YAFFS2, it can be concluded that every object header corresponds to a version of file. Once a transaction occurs, there will be an object header and a new version is created. So, the deleted file can be recovered until the respective block gets garbage collected. And versions of each database can be restored as much as possible.

4 The Proposed Method

The process framework of the proposed algorithm is shown in Fig. 1.

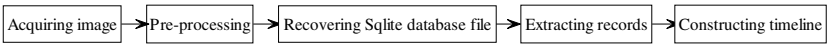


Fig. 1. The process framework of the proposed algorithm

There are two ways to acquire Android image—physical and logical method. Physical method is carried by JTAG [8] while logical method is carried out by "DD" or "NANDdump" instruction after rooting is executed. In this paper, only the logical method is considered.

The pre-processing sequentially records each chunk's objectID, objectType, chunkID, and chunkType in accordance with the allocation order on chip. All the blocks of a flash chip are sorted by *sequence number* from the largest to the smallest. And then

these sorted blocks are scanned from the one with the largest *sequence number* to the one with the smallest, and within a block, its chunks are scanned from the last one to the first. During this process, *objectID*, *objectType*, *chunkID* and *chunkType* of each chunk are stored into an array csq[] separately. A simplified algorithm of recovering sqlite database file in pseudo-code is shown below.

```
Algorithm: Recovering SQLite database file
Input: the array of structures csq[] and the image
Output: all SQLite files group by the filename
1.for each chunk in csq[] do /*scan the entire image reversely
     in chronological order*/
2.  if (chunkType=0x80) and (objectType=0x10) then
      /*recognize a file's object header chunk*/
3.    Find the data chunk of chunkID=1;
4.    if (magic number="SQLite format 3") then
        /*Read the magic number in file header and de-
        termine sqlite database file*/
5.      if the folder named by database's filename does not
          exist then
6.        create this folder;
7.      Extract information such as timestamps, objectID and
        file's length and store it;
8.      Calculate the num_chunks of the file by formula (1);
9.      Find all data chunks for the file;
10.     Calculate all data chunks' physical address by formula
        (2~4);
11.     Reassemble this file named by filename and its
        objectID in the folder;
```

$$num\_chunks = \lceil length / size\_chunk \rceil \tag{1}$$

Here *num_chunks* is the number of chunks the file occupies, *length* is the file's length, and *size_chunk* is the size of a page on NAND flash chip.

$$block\_offset = bsq[k/N] \tag{2}$$

$$chunk\_offset = N - (k\%N) \tag{3}$$

$$chunk\_address = block\_offset + chunk\_offset \tag{4}$$

Here k means the k^{th} chunk which was stored, N is the number of chunks in a block, *block_offset* is the chunk's physical block offset, *chunk_offset* is the chunk's relative offset in a block, and *chunk_address* is its physical address.

When scanning the whole chip, all sqlite database file are recovered indicated by object header. During this process, all files are grouped by the filename, and then distinguished by the history version number. In next step, all the records are extracted from each recovered integrated database file and stored into a CSV file, and verify that the integrity of the file that we restored.

When all recoverable history version of a same database file are recovered, a timeline for SQLite CRUD operations can be constructed by timestamp recorded in object header. In contrast of the SQLite records extracted from the two adjacent versions files, the SQL event from one change to another can be inferred. Then, through the analysis of the low-level SQL events corresponding to each user action, a user actions timeline can be constructed by analyzing the entire timeline of SQL event. Finally, we can have a global awareness about what the user did and when.

5 Experiments

In this part, a publicly dataset experiment is used to verify the effectiveness of the recovery method we proposed in the real scene.

The DFRWS has created two scenarios for the forensics challenge in 2011[9]. Images for Scenario 2 were acquired through NANDdump that OOB area can be acquired. 269MB File mtd8.dd for Scenario 2 was used in this experiment because it is the user image that contains a large number of user information. 110 sqlite files of different versions are recovered. Comparing our experimental results with the DFRC team's result, the latest version of all files in our result is equal to the DFRC team's result. In addition, the older versions of the sqlite file can be recovered using our method and the user actions can be analyzed using these files.

For example, the recovered file mmssms.db contains the user's short messages sent and received using this device. In our result, 16 versions file are recovered. Then a timeline **Fig. 2** and **Fig. 3** can be constructed according to these files. In the **Fig. 3**, you can clearly see what the user did and when. For instance, insert a record of id=17 in 04:43:04 represents the user received a short message. Update a record in 04:43:13 represents the user read it. Similarly, other database files can be analyzed too. Through a joint analysis of the results of all databases, a whole timeline of user behavior can be obtained. Experiment in this part shows that our proposed method is suitable for the real case and play an important role in forensic work.

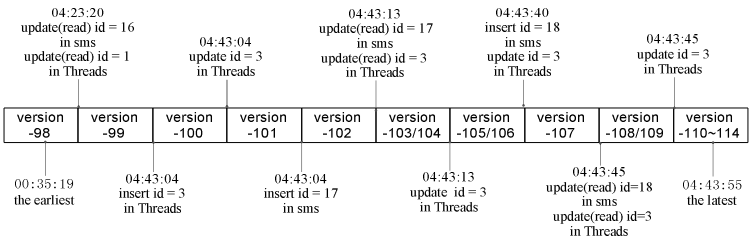


Fig. 2. Timeline of the SQL event about mmssms.db

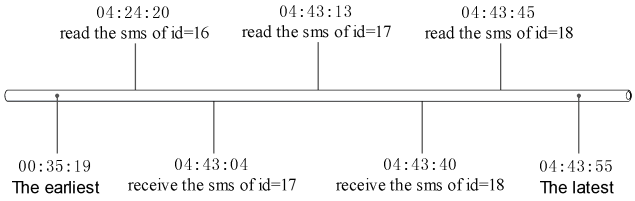


Fig. 3. Timeline of the user action about mmssms.db

6 Conclusions and Future Work

In this paper, a recovery method for SQLite record based YAFFS2 is proposed. Then construct timeline for SQLite CRUD operations by utilizing timestamp recorded in object header, and these may supply significant information about user behaviors to forensic investigations. The experimental results show the efficiency of the proposed method. This paper proves that file recovering from flash chips is practical, but as ext4 file system is widely used in android phone and Linux system, more forensic research is needed on ext4 in digital forensic perspective. Thus the technology of recovering data records from SQLite in ext4 file system will be our research direction.

Acknowledgements. This work is supported by the NSF of China under Grant No. 61070212 and 61003195, the Zhejiang Province NSF under Grant No. Y1090114 and LY12F02006, the Zhejiang Province key industrial projects in the priority themes under Grant No 2010C11050, and the science and technology search planned projects of Zhejiang Province (No. 2012C21040).

References

1. Quick, D., Alzaabi, M.: Forensic Analysis of the Android File System Yaffs2. In: 9th Australian Digital Forensics Conference (2011)
2. Manning, C.: How YAFFS works (2012), http://www.yaffs.net/documents/how-yaffs-works
3. Manning, C.: YAFFS Spec (2012), http://www.yaffs.net/yaffs-2-specification
4. Haerder, T., Reuter, A.: Principles of transaction-oriented database recovery. ACM Comput. Surv. 15(4), 287–317 (1983)
5. Pereira, M.T.: Forensic analysis of the Firefox 3 Internet history and recovery of deleted SQLite records. Digital Investigation 5(3), 93–103 (2009)
6. Jeon, S., Bang, J., Byun, K., et al.: A Recovery Method of Deleted Record for SQLite Database. Pers. Ubiquit. Comput. 16(6), 707–715 (2012)
7. Atomic Commit in SQLite (2012), http://www.sqlite.org/atomiccommit.html
8. Breeuwsma, M.: Forensic imaging of embedded system using JTAG(boundary-scan). Digital Investigation (2006)
9. DFRWS. DFRWS-2011-challenge (2011), http://www.dfrws.org/2011/challenge/index.shtml

UVHM: Model Checking Based Formal Analysis Scheme for Hypervisors

Yuchao She[1,*], Hui Li[1], and Hui Zhu[1,2]

[1] State Key Laboratory of Integrated Service Networks (ISN),
Xidian University, Xi'an 710071, P.R. China
sheyuchao@gmail.com
[2] Network and Data Security Key Laboratory of Sichuan Province,
Chengdu 611731, P.R. China

Abstract. Hypervisors act a central role in virtualization for cloud computing. However, current security solutions, such as installing IDS model on hypervisors to detect known and unknown attacks, can not be applied well to the virtualized environments. Whats more, people have not raised enough concern about vulnerabilities of hypervisors themselves. Existing works mainly focusing on hypervisors' code analysis can only verify the correctness, rather than security, or only be suitable for open-source hypervisors. In this paper, we design a binary analysis tool using formal methods to discover vulnerabilities of hypervisors. In the scheme, Z notation, VDM, B, Object-Z or CSP formalism can be utilized as suitable modeling and specification languages. Our proposal sequently follows the process of disassembly, modeling, specification, and verification. Finally, the effectiveness of the method is demonstrated by detecting the vulnerability of Xen-3.3.0 in which a bug is added.

Keywords: hypervisor, security, model checking, formal analysis.

1 Introduction

Cloud computing is a significant technology at present. The software that controls virtualization is termed as a hypervisor or a virtual machine monitor (VMM) that is seen as an efficient solution for optimum use of hardware, improved reliability and security.

Although there are many benefits, cloud computing encounters critical issues of security and privacy. Hypervisors have already become the path of least resistance for one guest operating system to attack another and it is also the path of least resistance for an intruder on one network to gain access to another network. The most important security issues for hypervisors are typically the risk of information leakage caused by information flow security weakness, etc. Some vulnerabilities of hypervisors have already been reported[1][2].

* Corresponding author.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 300–305, 2013.
IFIP International Federation for Information Processing 2013

Our Contribution. In this paper, we propose UVHM to detect vulnerabilities of hypervisors. In order to find as many vulnerabilities in the hypervisors as possible, the evaluation process must include demonstration of correct correspondences between security policy objectives, security specifications, and program implementation. Thus, we could use model checking theory[3][4] to discover vulnerabilities.

Related Work. Vulnerability analysis on hypervisors basically remains as a challenge. There are some existing works heavily focusing on code verification and hypervisor analysis. VCC[5] focuses on verifying the correctness of software rather than the security of it. Moreover, it can only verify C language. The Xenon project[6] is only suitable for open-source hypervisors. For Maude[7], the algebraic specification-based approach does not apply to analyze the vulnerabilities of VMMs. The existing models have a lot of limitations and can not pretend to address all of the security requirements of a system. Most of the available model checkers[8][9] use a proprietary input model. In summary, new studies have to be carried out basically starting from scratch.

2 Formal Analysis on Hypervisors

In UVHM, we develop suitable formal models, verification tools and related security policies according to our own needs to conduct more comprehensive studies on different aspects of hypervisor's security. Practical hypervisors' different design, architectures and working mechanisms will lead to different models, security policies, etc.

2.1 Formal Analysis on Binary Code

The scheme follows the process of disassembling – modeling – specification – verification. The general flow chart of UVHM is shown below.

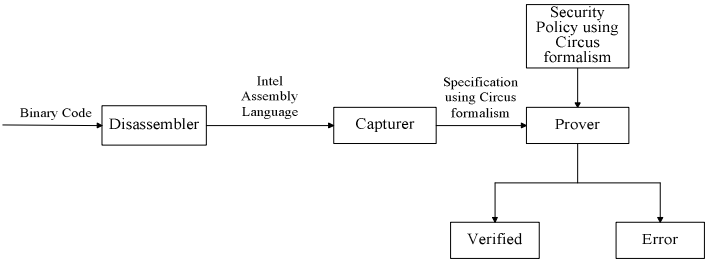


Fig. 1. The Flow Chart of UVHM

We shall first disassemble the hypervisor's binary file, and then formally model definitions of security, capture the behaviours of hypervisor's interfaces with such formal model, and verify the security using self-developed prover under the verification conditions.

1) Disassembling

We present static analysis techniques to correctly disassemble binaries and use at least two different disassemblers. The latter disassembler shall help fulfil some special requests/cases which cannot be handled by the former.

2) System Modeling

1. The self-developed formal models are needed. This model should contain the following characteristics: accurate, unambiguous, simple, abstract, easy to understand and only related with security. Only related with security means that the models only pay attention to the security features, and will not involve too many about functions and details of the implementation.
2. A great many hypervisors need hardware-assistant virtualization. Thus, we could adopt Z notation, VDM, B, Object-Z or CSP formalism to analyze concurrent process, and choose these formalism to define security. The partial orders of the system can be modelled into a lattice[10]. The most important relationship to be captured is probably the triangular dependency between three major entities from the state space: virtual contexts for guest domains, virtual instruction set processor VCPUs, and virtual interrupts or event channels. The mutual dependence between key components is a common feature in kernel design.

3) Specification

Unambiguous, precise specification of our requirement is needed. Integrality of security policy's specifications need to pay attention to. We could define some special hypercall interface sequences in security policy to identify illegal codes which execute in either guest or host domain and attempt to access another domain without permission.

For inter-domain security infringement, covert channel analysis will be adopted. Meta-flows[11] are combined to construct potential covert channels. Figure 2 shows the scene that the extension of f to mf is supervised by a series of rules. In this framework, we should define illegal flows in the form of information flow sequence, i.e., define the flow security policy.

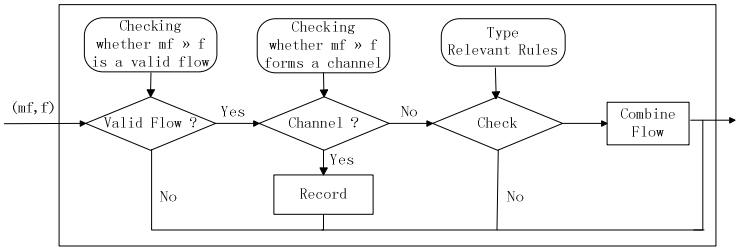


Fig. 2. Framework of Covert Channel Identification

4) Verification

Automated verification of a representative subset will be able to provide some critical insights into the potential difficulties and reveal the approaches that should be avoided.

3 System Implementation and Testing

We choose Xen-3.3.0 as our experimental subject. We use UVHM to verify whether the Xen contains the bug numbered 1492 in Xen's official website.

Before disassembling, we add this bug to Xen and compile it into hypervisor's binary file. Then, we use UVHM to get the whole formal analysis tool.

3.1 Adding the Bug

Add "free(buf); buf=NULL" to the file "tools/python/xen/lowlevel/acm/acm.c". Xen with the bug above could not detect the installed DEFAULT policy and reports the DEFAULT policy as "None" after initializing XSM-ACM module successfully.

There are two pictures to make a comparison between the installed Xen with the bug and without it.

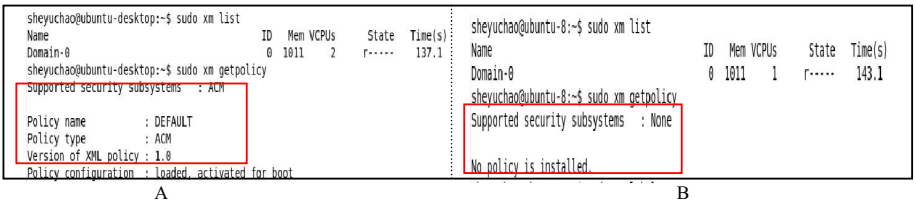


Fig. 3. Comparison Picture

Figure 3A shows that the DEFAULT security policy in the secure Xen is ACM whose version is 1.0, and it could be used as normal. Figure 3B shows that for the vulnerability added Xen, the DEFAULT security policy could not be used.

3.2 Implementation Module

1) Disassembling

We use IDA Pro, and BitBlaze to disassemble acm.o file. We could build up our models through analyzing the assembly language they gives us.

2) Modeling

What we concern about is whether the buffer where ACM policy loaded in is 'NULL' after the XSM-ACM module was initialized successfully.

Only several states that related with the buffer's state are being defined. We don't capture assignment instructions' behaviors which appeared in the assembly code which have nothing to do with the buffer's state.

3) Specification

If the buffer is 'NULL', of course, there is no policy could be used. We define this situation as a vulnerability. If not, the bug which the Xen contains is not the one defined above. Thus, we can define the following secure policy:

1) The buffer is 'NULL': This is a vulnerability caused by some wrong operations to the buffer, flag = 1 ;

2) The buffer is not 'NULL': Success, flag = 0.

4) Verification

Combining the model and specification together, we can get the tool. The input variables and relations among these variables can be regarded as an initial state. Based on the different range of the variables, the branch conditions will send them to different states. We could judge whether this is the vulnerability we defined through detecting the value of the flag. The following chart shows the visible model of the assembly code.

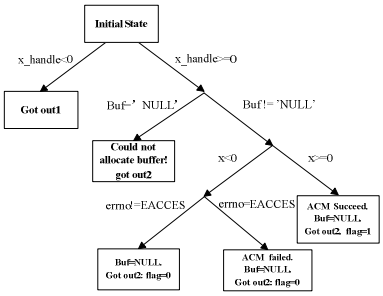


Fig. 4. The Visible Model

Now, the binary analysis tool is accomplished. We could use this tool to detect Xen hypervisors whether contains the vulnerability or not.

3.3 System Testing and Results Analysis

First, we disassemble the acm.o binary file. According to the assembly code and the defined model, we then sequently input needed variables or relations between them after analyzing the semantics of its assembly code.

1) For Xen with the bug, we input the following information after analysis: x_handle=32, x_op =1, buf != NULL, errno != EACCES. The system's report tells us this Xen contains the vulnerability we defined in the secure policy.

2) For Xen without the bug, we input the information: x_handle=6, x_op=-9, buf != NULL , errno != EACCES. The report says this Xen doesn't contain the vulnerability we defined.

Thus, without installing Xen, we are able to kown whether the Xen contains this bug.

This demonstrates the effectiveness of our formal binary analysis framework. The model and specification are all written in C language. They are linked through the flag.

4 Conclusion

There are security challenges in the cloud, and a secure cloud is impossible unless the virtual environment is secure. Aiming at this problem, we present our formal method which follows the process of disassembling – modeling – specification – verification to analyze the vulnerabilities of various hypervisors, etc.

We use this idea to realize a system that could verify whether the Xen contains the bug that will prevent the ACM policy from being used although the XSM-ACM module has been initialized successfully through analyzing its binary code. This demonstrates the effectiveness of the above method. This approach can be applied to detect vulnerabilities of various kinds of hypervisors.

References

1. Marshall, D.: Microsoft Hyper-V gets its first security patch. Infoworld (February 2010), http://www.infoworld.com/d/virtualization/microsoft-hyper-v-gets-its-first-security-patch-106
2. Vulnerability report: MS11-047 – Vulnerability in Microsoft Hyper-V could cause denial of service (June 2011), http://www.sophos.com/support/knowledgebase/article/113734.html
3. Clarke, E., Grumberg, O., Long, D.: Model Checking. MIT Press (1999)
4. Baier, C., Katoen, J.P.: Principles of Model Checking. MIT Press (2008)
5. Leinenbach, D., Santen, T.: Verifying the Microsoft Hyper-V Hypervisor with VCC. In: Cavalcanti, A., Dams, D.R. (eds.) FM 2009. LNCS, vol. 5850, pp. 806–809. Springer, Heidelberg (2009)
6. Freitas, L., McDermott, J.: Formal methods for security in the Xenon hypervisor. International Journal on Software Tools for Technology Transfer 13(5), 463–489 (2011)
7. Webster, M., Malcolm, G.: Detection of metamorphic and virtualization-based malware using algebraic specifi cation. In: EICAR 2008 (2008)
8. Holzmann, G.J.: The SPIN Model Checker: Primer and Reference Manual. Addison-Wesley (2004)
9. Ball, T., Levin, V., Rajamani, S.K.: A Decade of Software Model Checking with SLAM. Communications of the ACM 54(7), 68–76 (2011)
10. Denning, D.: lattice model of secure information flow. Communications of the ACM 19(5), 236–243 (1976)
11. Shen, J., Qing, S.: A Dynamic Information Flow Model of Secure Systems. In: CCS, pp. 341–343 (2007)

SA4WSs: A Security Architecture for Web Services

Lingxia Liu[1,2,3], Dongxia Wang[1,2], Jinjing Zhao[1,2], and Minhuan Huang[1,2]

[1] Beijing Institute of System Engineering, Beijing, China
[2] National Key Laboratory of Science and Technology on Information System Security, Beijing, China
[3] The Information and Navigation Institute, Air Force Engineering University, Xi'an, China
lingxia_liu@tom.com, dongxiawang@126.com, misszhaojinjing@hotmail.com, huangmh06@mails.tsinghua.edu.cn

Abstract. With the rapid development and wide application of the Web services, its security flaws and vulnerabilities are increasing. Security has become one of the key issues to constrain the development of Web services technology. In this paper, we focus on how to build a security architecture for Web services to meet the security requirements of Web service applications. On the basis of analyzing the existing methods, a new security implementation approach for Web services is proposed to meet both the common security requirements of Web services platform and the specific security requirements of Web service applications. Then a security architecture for Web services is proposed. The architecture supports separating the functional implementations of Web service from the non-functional implementation of Web service, and ensures the portability of the platform.

Keywords: Web service, Security, Architecture.

1 Introduction

As a new Web application model, Web service is heterogeneous, dynamic and loose coupling, and introduces a great deal of special threats to Web service applications, and makes that traditional security techniques are inadequate to solve the security issues in Web services.

In recent years, many research institutions, organizations and companies devote to the research of security for Web Services. Organizations and research institutions for Standardization mainly concern in security standards for Web service. The academic community tries to solve the security issues from the theoretical aspect. Middleware companies and open source organizations try to provide security protection tools for customers by providing a set of software or toolkit.

In the paper, we focus on how to build security architecture for Web services to meet the security requirements of Web service applications. On the basis of the improved approach, a security architecture for Web services named SA4WSs (Security Architecture for Web Services) is proposed. The architecture supports separating function implementation of service from non-functional implementation of service, meanwhile ensure the portability of the platform.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 306–311, 2013.
IFIP International Federation for Information Processing 2013

The remainder of this paper is organized as follows. In section 2, the related works are reviewed. In section 3, the security architecture for Web services is set forth. Finally, in Section 4, we conclude the paper.

2 Related Works

Many Organizations for Standardization, research institutions and companies research on the security architecture for Web services.

2.1 Security Standards

In order to assure the end-to-end security for SOAP message and improve the interoperability of Web service, OASIS and other organizations define the Web Service Security model consisting of multiple standards [1]. The model is insufficient. It is only to protect the two trust sides to communication by a secure connection, and not address the security problem caused by anonymous consumers invoking Web service or SOAP API.

W3C proposed a standard named XML Signature Syntax and Processing specification, which defines how to sign part or all of one XML document [2]. W3C also proposed a standard named XML Encryption Syntax and Processing, which defines how to encrypt part or all of one XML document [3].

2.2 Security Architecture

U.S. DISA (Defense Information Systems Agency) released a specification named "Network Centric Enterprise Services security architecture (version 0.3)" in 2004, provides a service-oriented information security reference architecture to ensure the security of services in network-centric environment [4].

IBM proposes the service-oriented security reference modeling and architecture based on its own SOA infrastructure and business scenarios to cope with information security issues for work flows [5]. The service-oriented security architecture proposed by IBM is meaningful, but its portability is restricted because it relies heavily on its own SOA technologies and products.

GSI provides basic security services for grid computing environments. It is an integrated solution to solve the security issues in grid computing and becomes a standard of GGF [6]. The major security functions in GSI include Certificates, Mutual Authentication, Confidential Communication, Securing Private Keys, and Delegation and Single Sign-On. GSI also has some disadvantages, such as frequent, complex and poor scalable authentication between entities.

OGSA is proposed by Ian Foster et al. on the basis of five-level hourglass structure and Web services [7]. A basic premise of OGSA is that everything is represented by a service, not excepting security. All kinds of security mechanisms such as encryption, access control, and audit are represented as services to facilitate the implementation of the security-related policies.

The service oriented security architecture named SOSIE presented by [8] is realizing the security functions into modular, stand-alone security services. The article [9] addresses the question of security mechanisms that are usually used and that can be used in Web services based SOA implementation from standardized as well as technical and implementation point of view, and gives an overview of SOA security solutions. The article [10] provides a set of software architecture viewpoints that allow security architects to construct a holistic system design based on a set of views. Other related works focus on the special issues of security service [11] [12] [13].

2.3 Supporting Platform

In addition, some research institutions and companies have also developed multiple security platforms and middlewares for Web service.

Microsoft proposed a solution for Web services security problems named WSE (Web Services Enhancements) on the basis of .NET platform [14]. WSE is a service security middleware for .NET platform, including authentication and encryption library. It allows developers to develop secure Web service by implementing the latest WS-Security specification.

Globus Toolkit is the reference implementation of OGSI. It provides basic security services required by grid computing, including message protection, authentication, authorization, and audit/log [15].

The Apache WSS4J project provides a Java implementation of the primary security standard for Web Services, namely WS-Security [16]. WSS4J ships with handlers that can be used in Axis-based web services for an easy integration.

3 Security Architecture for Web Services

3.1 Improved Security Implementation Approach

Web service applications face threats including the threats the Web services platform faced where service located and the threats the service implementation faced, i.e. common threats the platform faced and specific threats the application faced. The first approach [8] can cope with specific threats the application faced, but cannot cope with the common threats the platform faced. The second and the third approaches can cope with the common threats the platform faced, but cannot cope with the specific threats the application faced.

We try to improve the three approaches to meet the security requirements of Web service applications. The first approach cannot cope with a lot of attacks (such as SOAP message replay attacks) the platform faced, no matter how to transform it, because the communication between service implementation and resource gateway is no longer via SOAP message. The third approach is similar to the firewall, and it is not suitable to add application-specific security mechanisms to cope with specific threats the application faced.

Therefore, we focus on improving the second approach to cope with both common threats the platform faced and specific threats the application faced. Analysis on the

Web services platform showed that the platform can be logically divided into two parts. One part of it is Web services runtime environment, mainly providing the functions including deployment, management, publishing and finding for Web service. The other part is resource gateway, which are a set of interfaces between Web services runtime environment and service implementation. It provides the protocol conversion functions for specific application-related invoking. In view of the above analysis, an improved security implementation approach is proposed, as shown in Figure 1.

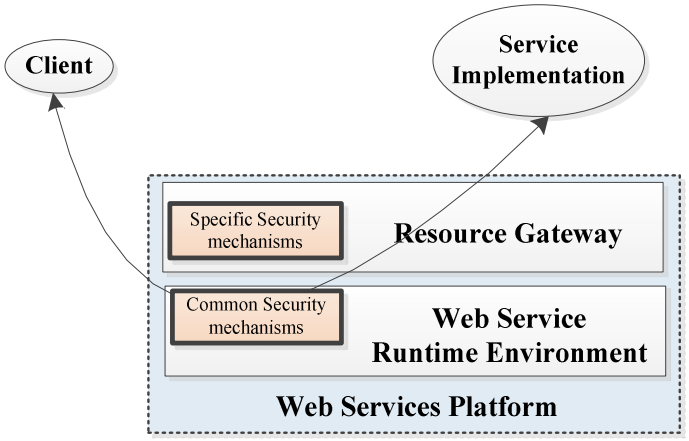


Fig. 1. The improved security implementation approach

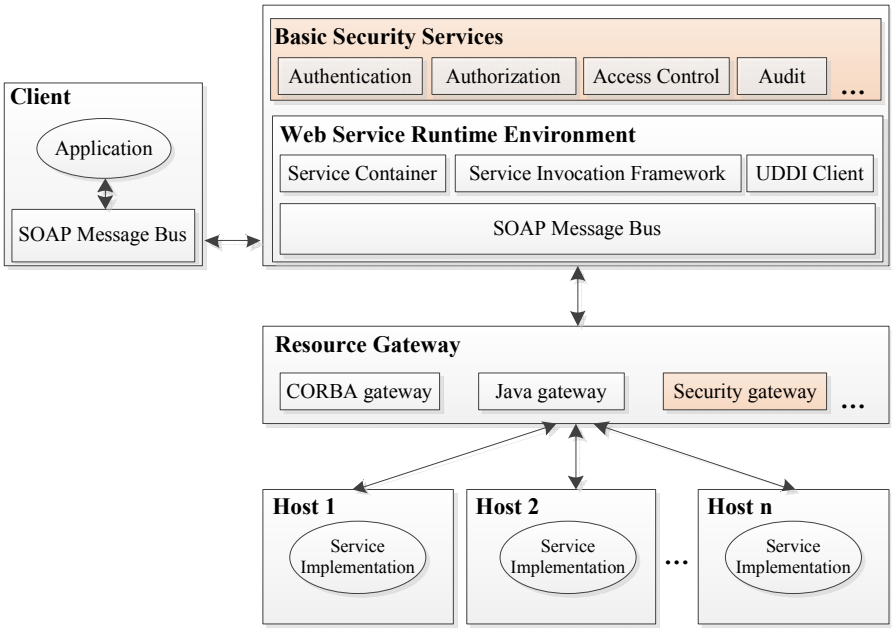


Fig. 2. Security architecture for Web services

In this approach, the security mechanisms are divided into two parts. One part of it is common security mechanisms, which is not relevant to applications and resides in the Web services runtime environment. The other part is specific security mechanisms, which is relevant to applications and resides in the resource gateway. The improved approach not only remains separating functional implementation from non-functional implementation, but also assures the portability of the platform.

3.2 Architecture

A security architecture for Web services named SA4WSs is proposed, as shown in Figure 2.

In the architecture, the module of Basic Security Services provides basic services for implementing the basic security functions, including authentication, authorization, access control and audit, etc. [8]. The module of Security gateway is responsible for the application-specific security functions, and the service implementation is responsible for the business logic. The Security gateway interacts with the service implementation via various protocols (The protocol type is decided by the type of service implementation).

4 Conclusion

In this paper, a security architecture for Web services is proposed. The architecture described in the paper supports separating the functional implementations from the non-functional implementation, and ensures the portability of the platform. SA4WSs has been partially implemented based on Apache Axis. Future work consists of implementing a Web Services security supporting platform to support the development and management of security Web service.

Acknowledgements. This research is supported by National Natural Science Foundation of China (Grant No. 61100223 and No. 61271252).

References

1. Gerié, S., Hutinski, Ž.: Standard Based Service-Oriented Security. In: 18th International Conference on Information and Intelligent Systems, pp. 327–335. IEEE Press, Croatia (2007)
2. W3C: XML Signature Syntax and Processing Version 2.0. Standard, W3C (2012)
3. W3C: XML Encryption Syntax and Processing Version 1.1. Standard, W3C (2012)
4. Defense Information Systems Agency. A Security Architecture for Net-Centric Enterprise Services (NCES) Version 0.3. Technical report, Defense Information Systems Agency (2004)
5. Buecker, A., Ashley, P., Borrett, M., Lu, M., Muppidi, S., Readshaw, N.: Understanding SOA Security Design and Implementation. Redbook, IBM (2007)

6. Overview of the Grid Security Infrastructure, http://www.globus.org/security/overview.html
7. Foster, I., Kishimoto, H., Savva, A.: The Open Grid Services Architecture, Version 1.5. Technical report, Global Grid Forum (2006)
8. Opincaru, C., Gheorghe, G.: Service Oriented Security Architecture. Enterprise Modelling and Information Systems Architectures 4(1), 39–48 (2009)
9. Gerić, S.: Security of Web Services based Service-oriented Architectures. In: MIPRO 2010, pp. 1250–1255. IEEE Press, Croatia (2010)
10. Peterson, G.: Service Oriented Security Architecture. Information Security Bulletin. 10, 325–330 (2005)
11. Lee, S.M., Kim, D.S., Park, J.S.: A Survey and Taxonomy of Lightweight Intrusion Detection Systems. Journal of Internet Services and Information Security 2(1/2) (February 2012)
12. Hori, Y., Claycomb, W., Yim, K.: Guest Editorial: Frontiers in Insider Threats and Data Leakage Prevention. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, JoWUA 3(1/2) (March 2012)
13. Ho, S.M., Lee, H.: A Thief among Us: The Use of Finite-State Machines to Dissect Insider Threat in Cloud Communications. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, JoWUA 3(1/2) (March 2012)
14. Web Service Enhancement 3.0, http://msdn.microsoft.com/en-us/library/bb896679.aspx
15. The Globus Alliance, http://www.globus.org/toolkit
16. Apache WSS4J, http://ws.apache.org/wss4j/

Verifying Data Authenticity and Integrity in Server-Aided Confidential Forensic Investigation

Shuhui Hou[1], Ryoichi Sasaki[2], Tetsutaro Uehara[3], and Siuming Yiu[4]

[1] Department of Information and Computing Science, University of Science and Technology Beijing, China
[2] Graduate School of Science and Technology for Future Life, Tokyo Denki University, Japan
[3] Research Institute of Information Security, Wakayama, Japan
[4] Department of Computer Science, The University of Hong Kong, Hong Kong

Abstract. With the rapid development of cloud computing services, it is common to have a large server shared by many different users. As the shared server is involved in a criminal case, it is hard to clone a copy of data in forensic investigation due to the huge volume of data. Besides, those users irrelevant to the crime are not willing to disclose their private data for investigation. To solve these problems, Hou et al. presented a solution to let the server administrator (without knowing the investigation subject) cooperate with the investigator in performing forensic investigation. By using encrypted keyword(s) to search over encrypted data, they realized that the investigator can collect the necessary evidence while the private data of irrelevant users can be protected from disclosing. However, the authenticity and integrity of the collected evidence are not considered there. The authenticity and integrity are two fundamental requirements for the evidence admitted in court. So in this paper, we aim to prove the authenticity and integrity of the evidence collected by the existing work. Based on commutative encryption, we construct a blind signature and propose a "encryption-then-blind signature with designated verifier" scheme to tackle the problem.

Keywords: confidential forensic investigation, authenticity and integrity, commutative encryption, signcryption.

1 Introduction

With the rapid development of cloud computing technology, forensic investigation is becoming more and more difficult as crimes occur. The traditional technique disk cloning may be impossible to conduct for collecting evidence from a data center due to the massive volume of data and the distributed manner of storage device(s). Besides, it is common to have a large server shared by many different users in the cloud computing environment. The shared server stores not only suspicious data relevant to the crimes but also stores an enormous amount of data involving *sensitive* information that is totally irrelevant to the crimes. The users irrelevant to the crimes may not want to release their information

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 312–317, 2013.

for investigation especially as it involves confidential or privacy information. To improve the investigation efficiency and protect the privacy of irrelevant users, one strategy is to let the server administrator search, retrieve and hand only the relevant data to the investigator, where the administrator is supposed to be responsible for managing the data in a secure manner. Due to some special crimes, the investigator may not want the administrator to know what he is looking for. In short, it is indispensible to consider how to protect both confidentiality of investigation and privacy of irrelevant users in such forensic investigation. For simplicity of description, we refer to this problem as "server-aided confidential forensic investigation".

To solve the problem "server-aided confidential forensic investigation", Hou et al. [1,2] presented several solutions under the assumption that the server administrator is willing to cooperate with the investigator to search the relevant data. The detail of their solutions is as follows: (1) the investigator specifies a single keyword or multiple keywords based on the investigation subject, encrypts it or them with his public key and sends the encrypted keyword(s) to the administrator; (2) with the investigator's public key, the administrator encrypts all the data files stored on the server. Then, he uses the encrypted keyword(s) to search over encrypted data files, retrieves and sends only the relevant data (i.e., those encrypted files whose corresponding plaintext files contain the specified keyword(s)) to the investigator; (3) the investigator decrypts the relevant data with his private key and performs investigation only on such relevant data for capturing the criminal evidence. The irrelevant data (those files without containing the keyword(s)) will never be sent to the investigator, so can be protected from exposing to the investigator. By using encrypted keyword(s) to search over encrypted data, the administrator has no idea of what the investigator is looking for.

In the above solutions, the administrator is supposed to have responsibility for protecting the irrelevant data against disclosing, and at the same time he is prevented from learning the relevant data due to some special crimes. But without learning what the relevant data is, the administrator cannot judge if the relevant data is really *relevant* to the crimes and neither can check if the investigator obtained other irrelevant data from the server. Regarding this problem, the work [1,2] assume that the administrator can require the investigator to show what data is collected based on what keyword(s) when the relevant data is presented as evidence in court. However, even if the assumption works, no measures can guarantee that the presented data is the one that comes from the server and no alteration occurs to it. In other words, the authenticity and integrity of the evidence collected in the work [1,2] are not considered. The authenticity and integrity are two fundamental requirements for admissibility of evidence in court and they are crucial to win a case. Therefore, we put our major concern on how to prove the authenticity and integrity of the evidence collected in the work [1,2].

In this paper, we propose a "encryption-then-blind signature with designated verifier" scheme to prove the authenticity and integrity of the evidence. When

the above-mentioned relevant data is presented as evidence during a trial, we aim to realize that the administrator (or the third party the administrator trusts) can verify whether the presented evidence is the data that comes from the server and whether the evidence is altered or not. In addition, we implement the proposed system based on commutative encryption and examine its security.

2 Preliminaries: Commutative Encryption

In our proposed scheme, commutative encryption plays an important role.

Definition 1. *Let $\mathcal{M}$ denote a message space, $\mathcal{K}$ denote a key space and $\mathcal{C}$ denote a cipher message space, respectively. A commutative encryption function is a family of bijections $\mathcal{E}$: $\mathcal{M}\times\mathcal{K}\to\mathcal{C}$ such that for a given $m\in\mathcal{M}$ we have $\mathcal{E}_{k_1}(\mathcal{E}_{k_2}(m))=\mathcal{E}_{k_2}(\mathcal{E}_{k_1}(m))$, for any $k_1,k_2\in\mathcal{K}$.*

It follows that if a message is encrypted by two different keys k_1 and k_2, it can be recovered by decrypting the cipher message with k_1 followed by decrypting with k_2. The message can also be recovered by decrypting with k_2 followed by decrypting with k_1.

The RSA cryptosystem is commutative for keys with a common modulus n. It was introduced by Rivest, Shamir and Adleman in 1978 ([3]) and one description of the system is described below.

RSA Cryptosystem. Let $n=pq$ where p and q are a pair of large, random primes. Select e and d such that $ed=1 \pmod{\phi(n)}$ where $\phi(n)=(p-1)(q-1)$. n and e are public while p,q and d are private.
The encryption operation is:

$$c = ENCRYPT(m) = m^e \mod n$$

The decryption operation is:

$$m = DECRYPT(c) = c^d \mod n$$

Where m is the plaintext message and c is the resulting ciphertext.

Using $\mathcal{E}_k(\cdot)$ to denote the encryption operation with key k, it is obvious that

$$\begin{aligned}\mathcal{E}_{e_1}(\mathcal{E}_{e_2}(m)) &= (m^{e_2})^{e_1} \bmod n \\ &= m^{e_2 e_1} \bmod n \\ &= m^{e_1 e_2} \bmod n \\ &= (m^{e_1})^{e_2} \bmod n \\ &= \mathcal{E}_{e_2}(\mathcal{E}_{e_1}(m)) \bmod n\end{aligned}$$

i.e., the RSA cryptosystem is commutative for keys with a common modulus n.

3 Proposed Scheme: Encryption-Then-Blind Signature with Designated Verifier

3.1 Requirements of "Server-Aided Confidential Forensic Investigation"

For clarity, we summarize the requirements of "server-aided confidential forensic investigation" below.

From **investigator's** viewpoint, he hopes to fulfill the following:

- Collect evidence only from relevant data for saving time and effort, so as to improve the investigation efficiency;
- Let server administrator search and retrieve relevant data but without letting him know what he is searching and retrieving;
- Verify the authenticity and integrity of the relevant data so that it can be admitted in court when it is presented as evidence.

From **administrator's** viewpoint, he hopes to fulfill the following:

- Protect irrelevant data against exposing while cooperating with investigator in collecting evidence, i.e., ensuring that no privacy of irrelevant users leaks during investigation;
- Be able to verify the authenticity and integrity of the relevant data when it is open or presented as evidence in court. That is, the administrator needs to protect user data against unauthorized disclosing. If some data has to be open, it should be open in a secure manner.

3.2 Details of Proposed Scheme

For the brevity of description, we take single keyword case as an example and adopt the following notation. The single keyword specified by the investigator is denoted as w^*, which is l-bit long; The data stored on the server is assumed to be a set of documents, denoted as $\{W^1, W^2, \ldots, W^L\}$. A document $W \in \{W^1, W^2, \ldots, W^L\}$ consists of a sequence of words, denoted as $W=\{w_1, w_2, \ldots, w_v\}$ where every word w_i is l-bit long. We also assume that both w^* and W come from the same domain. It should be pointed out that a document does not always consist of equal-length words, but we can transform the variable-length words into fixed-length words through hashing. The encryption of w^* and W is denoted as $\mathcal{E}(w^*)$ and $\mathcal{E}(W)=\{\mathcal{E}(w_1), \mathcal{E}(w_2), \ldots, \mathcal{E}(w_v)\}$, where $\mathcal{E}(\cdot)$ is the encryption function.

Assume that there is a secure channel between server administrator and investigator. Based on commutative encryption, the "encryption-then-blind signature with designated verifier" scheme works as follows.

1. **Encryption** for confidentiality and privacy
   For the confidentiality of investigation, the investigator encrypts his specified keyword w^* with his public key p_I and sends the administrator the encrypted

keyword $\mathcal{E}_{p_I}(w^*)$ as well as his public key p_I; on server side, the administrator encrypts all the documents $\{W^1, W^2, \ldots, W^L\}$ with the public key p_I, where the resulting documents are denoted as $\{\mathcal{E}_{p_I}(W^1), \mathcal{E}_{p_I}(W^2), \ldots, \mathcal{E}_{p_I}(W^L)\}$. Both the keyword and the documents are encrypted, which are assumed to be provably secure in the sense that the administrator cannot learn anything about the specified keyword as it is encrypted and the investigator cannot learn more than the searching results. The searching results must contain the specified keyword, so the investigator can treat them as potential evidence.

2. **Blind signature** for authenticity and integrity
   On server side, the administrator performs the following.
   - uses $\mathcal{E}_{p_I}(w^*)$ to search over all the encrypted documents $\{\mathcal{E}_{p_I}(W^1), \mathcal{E}_{p_I}(W^2), \ldots, \mathcal{E}_{p_I}(W^L)\}$, and retrieves $\mathcal{E}_{p_I}(W)$ such that the plaintext document W contains the keyword w^* (i.e., $W \ni w^*$).
     There are several ways to judge whether the plaintext document W contains the keyword w^* based on the relation between the ciphertext $\mathcal{E}(W)$ and $\mathcal{E}(w^*)$. As $\mathcal{E}(\cdot)$ is a deterministic encryption (e.g., RSA cryptosystem), we can get $W \ni w^*$ if $\mathcal{E}(W) \ni \mathcal{E}(w^*)$, i.e., there exist one word $w_i \in W$ such that $\mathcal{E}(w_i) = \mathcal{E}(w^*)$; as $\mathcal{E}(\cdot)$ is a probabilistic encryption, $W \ni w^*$ can be shown by applying techniques like zero-knowledge proof (please refer to the work [1] where Paillier cryptosystem is used).
   - signs W blindly by computing $\mathcal{E}_A(\mathcal{E}_{p_I}(W))$ if $W \ni w^*$ and sends the investigator $\mathcal{E}_{p_I}(W)$ as well as $\mathcal{E}_A(\mathcal{E}_{p_I}(W))$, where $\mathcal{E}_A(\cdot)$ is commutative encryption with $\mathcal{E}_{p_I}(\cdot)$ and the subscript 'A' means that it is the administrator's encryption function. As $\mathcal{E}_A(\cdot)$ is public key encryption, $\mathcal{E}_A(\mathcal{E}_{p_I}(W))$ means encrypting $\mathcal{E}_{p_I}(W)$ with the public key of the administrator. As $\mathcal{E}_A(\cdot)$ is secret key encryption, the secret key is only known to the administrator. In the following, we consider that $\mathcal{E}_A(\cdot)$ is public key encryption. Here, the other documents without containing the keyword w^* will never be sent to the investigator, so their privacy can be protected completely.
     The signature has the following properties: **(a) Selective signature**: the administrator signs only the relevant data W ($W \ni w^*$) instead of all the data stored on the server for less computational cost; **(b) Blind signature**: the administrator wants to verify the authenticity and integrity of the original relevant data W ($W \ni w^*$) rather than its illegible encrypted form $\mathcal{E}_{p_I}(W)$, so he needs to sign W blindly, that is, sign W without knowing what the W is. Here, the administrator signs W blindly by computing $\mathcal{E}_A(\mathcal{E}_{p_I}(W))$, i.e., computing encryption of W twice; **(c) Designated verifier signature**: the administrator wants to check if the relevant data is really *relevant* to the crimes and ensure that the investigator does not obtain other irrelevant data from the server. The administrator needs to control the signature verification. On the other hand, the investigator also needs the administrator's cooperation to prove that the relevant data does come from the server and no alteration occurs to it when it is presented as evidence. In a word, a designated verifier signature rather than public verified signature is required here. In our

scheme, only the administrator knows signing key and verification key, so only the administrator can verify the signature. The administrator can also delegate the verification key to the third party he trusts and let the third party verify the signature.

3. **Decryption**
   The investigator decrypts $\mathcal{E}_{p_I}(W)$ ($W \ni w^*$) with his private key and performs investigation on W for capturing evidence. He also decrypts $\mathcal{E}_A(\mathcal{E}_{p_I}(W))$(which is $\mathcal{E}_{p_I}(\mathcal{E}_A(W))$ as $\mathcal{E}_{p_I}(\cdot)$ and $\mathcal{E}_A(\cdot)$ are commutative) for obtaining the signed W, i.e., $\mathcal{E}_A(W)$. He keeps the $\mathcal{E}_A(W)$ for the later signature verification.
4. **Signature verification**
   When the W is presented as evidence in court, the administrator (or the third party the administrator trusts) verifies the signature by test if $\mathcal{D}_A(\mathcal{E}_A(W))=W$ is true, where $\mathcal{D}_A(\cdot)$ is the inverse of encryption process $\mathcal{E}_A(\cdot)$. In other words, the administrator (or the third party the administrator trusts) verifies if the evidence is the W that comes from the server and if it is altered or not. At the same time, this also helps the investigator to show the authenticity and integrity of the evidence.

4 Conclusions

The scheme proposed in the paper can be shown to satisfy all security requirements. We also implemented our scheme based on an RSA cryptosystem. The results show that the performance is acceptable. Due to the space limitation, both the details of the security analysis and the experimental results will be given in the full paper. For future work, we will consider multi-dimensional search (e.g., range search, equality search, etc.) over encrypted data to overcome the restriction of the keyword search.

Acknowledgments. This work is partially supported by "Heiwa Nakajima Foundation, Japan" and partially sponsored by "the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry".

References

1. Hou, S., Uehara, T., Yiu, S.M., Hui, L.C.K., Chow, K.P.: Privacy Preserving Confidential Forensic Investigation for Shared or Remote Servers. In: 2011 Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 378–383 (2011)
2. Hou, S., Uehara, T., Yiu, S.M., Hui, L.C.K., Chow, K.P.: Privacy Preserving Multiple Keyword Search for Confidential Investigation of Remote Forensics. In: 2011 Third International Conference on Multimedia Information Networking and Security, pp. 595–599 (2011)
3. Rivest, R.L., Shamir, A., Adleman, L.: A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. Communications of the ACM 2(21), 120–126 (1978)

A Test Case Generation Technique for VMM Fuzzing

Xiaoxia Sun[1,2], Hua Chen[1,2], Jinjing Zhao[1,2], and Minhuan Huang[1,2]

[1] Beijing Institute of System Engineering
[2] National Key Laboratory of Science and Technology on Information System Security
xiaoxia83@gmail.com,
{chenhua8011,misszhaojinjing}@hotmail.com,
huangmh06@mails.tsinghua.edu.cn

Abstract. In this paper, we first give a short introduction to the security situation of virtualization technology, and then analyze the implementation challenges of the CPU virtualization component of a hybrid system virtual machine with support of running a guest machine of the IA-32 instruction set. Based on a formal definition of the CPU's execution state, we propose a fuzzing test case generation technique for both the operands and operators of instructions, which can be applied to fuzz testing the virtual machine monitor (VMM) of a hybrid system virtual machine.

Keywords: VMM, fuzzing, IA-32.

1 Introduction

In the name of cost savings, more and more critical systems and business systems have been virtualized. Growing interests in cloud computing will fuel further demand on virtualization. Thus the attentions on virtualization system security, specifically on business virtualization system have grown as interest has grown.

As a most-widely used virtualization technology, system virtual machines make it possible that multiple OS environments (guest machine) coexist on the same physical computer (host machine), and the ISA (instruction set architecture) of guest machines can be different from the host machines. This technology not only provides huge saving of hardware costs, but also brings many security advantages. Virtualization can offer isolated execution environment make services of different security policies running in separate virtual machines which is actually the same computer, without any interference of each other.

Virtual Machine (VM) has become a new battle of security attacks and defenses [1, 2, 3]. More and more malwares and attacks turn to choose VM as their targets, and in recent years, many traditional defense techniques have their corresponding counterpart of VM platforms, such as VM-based intrusion detection system, intrusion prevention system and honeypot. However, the security of virtual machine [4] itself is still a problem to solve. On the other hand, the business system has high requirements on security. The system needs to give the right users access to the right resources at the

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 318–323, 2013.
IFIP International Federation for Information Processing 2013

right time, protect sensitive business data, keep applications available and keep away from malicious or fraudulent use.

An increasing number of virtualization system vulnerabilities have been reported. As reported by the IBM X-force [5], nearly one hundred vulnerabilities on virtualization system have been disclosed every year since 2007. And among 40% of them have high severity, which is easy to be exploited, and provide full control over attacked systems due to the intrinsically high privilege level of virtual machine monitor (VMM) itself. Among all these vulnerabilities, the VMM vulnerabilities and VMM escape vulnerabilities are of the most severe, since they will compromise all guest VMs. For instance, by modifying the processor status register, a local attacker can cause the Xen kernel to crash [6], and an error in the virtual machine display function on VMware ESX Server allows an attacker in a guest VM to execute arbitrary code in the VMM [7]. More and more attention has been paid to VM security, especially VMM security.

The rest of the paper is organized as follows: Section 2 discusses implementation challenges of IA-32 ISA based VMM. Section 3 introduces the test case generation technique for VMM fuzzing. Section 4 draws the conclusions.

2 Implementation Challenges of IA-32 ISA Based VMM

To provide an isolated execution environment for guest OS, the system VM needs to interpret, translate and emulate the Instruction Set Architecture (ISA) of guest machine. The guest ISA in this paper is the pervasive Intel IA-32 instruction set, a Complex Instruction Set Computing (CISC). Compared with the Reduced Instruction Set Computing (RISC), IA-32 is more complex in instruction type (e.g. different kinds of memory access instructions exist), and complicated instruction encoding method .The encoding length of IA-32 is not fixed and there is redundancy at the opcode level. Instruction prefixes can further complicate the instruction semantics. All of these lead make it hard to emulate the IA-32 instructions. Therefore, IA-32 cannot be virtualized efficiently. For example, to guarantee the security of the virtualization system, some special handle needs to be employed to some instruction of IA-32 ISA.

The inherent complexity of IA-32 instructions makes it very hard to precisely model the semantic behavior of IA-32 instructions. In modern processers, there are two types of instructions according to execution mode, privileged instructions and non-privileged instructions. The privileged instructions only can be executed in the system mode, and when executed in the user mode, it will cause a trap. However, there are some non-privileged instructions in IA-32, may access some sensitive resources in the CPU, and modify the resource configuration of the system. We call these sensitive instructions as critical instructions [8]. Robin and Irvine [9] have figured out there is seventeen critical instructions that cannot be efficiently virtualized, which can be divided into two categories: the system-protection instructions and register-sensitive instructions. The system-protection instructions may access and modify the memory allocation system, including LAR, LSL, VERR, VERW, POP, PUSH, CALL, JMP, INT n, RET STR and MOVE. And the register-sensitive instructions

may access and modify the configuration or CPU state register, including SGDT, SIDT, SLDT, SMSW, PUSHF and POPF. These instructions may have side effect, that is to say, the execution of an individual instruction may alter values of CPU status register (*e.g.* EFLAGS), thus may have an effect on the behavior of other instruction implicitly.

To guarantee the security of the whole VM system, VMM have to introduce a so-called scan and patch mechanism to detect such critical instructions and when execute these instructions, trigger a trap instead. Then the VMM handle the trap by emulating these instructions properly.

It is hard to prove a VMM is security in formal, so we need tests to validate the security of VMM. It was proved that fuzzing [10] is a useful test method for many systems, especially when input space is rather large. In this paper, we employ the fuzzing for validating the security of VMM.

From above discussion, it is prone when executing the critical instructions on VMM, so we can conclude that the key aspect to test whether a VMM is security is test the execution of the critical instructions. In this paper, we focus on testing such critical instructions. In next section, we will introduce the proposed test case generation technique on VMM fuzzing.

3 Test Case Generation and Execution

3.1 Test Procedure Overview

The machine can exist in any one of a finite number of states where each state has four components, we represent the state of the abstract machine with the tuple s = (PC; R; M; E). The registers state R is a mapping from registers to their values stored there. The memory M is a mapping from memory addresses to values the particular location store. M store the instructions and R store data, which is similar to text section and data section. The program counter PC can refer to any memory address where stored the next instruction CPU is to execute; E is to denote the termination of the execution or exception. In this way, the CPU can be regarded as a finite state machine.

Virtualization transparency characteristics mean that one cannot tell whether a program is executed in physical machine or a virtualized one. According to the transition system definition given above, a virtual machine is transparent if and only if for any possible equivalent beginning states pairs (denoted as S_b and S_b'), that is, after arbitrary legal instructions executed simultaneously on the physical machine and virtual machine from same beginning state S_b, the states both machines reach (denoted as S_e and S_e') are semantically equivalence. By feeding testcases mutating the type of instructions, prefixes and operand instructions, we can find a crash report or get both final states and then compare them to find differences. A difference always indicates a virtualized machine have a violation of Virtualization transparency, which means the final states are not same with same beginning state and same operation. Fig.1 shows the overview of the test procedure.

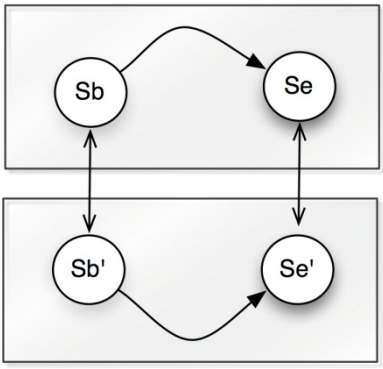


Fig. 1. The overview of test procedure

```
void main(){
    void *p;
    char *code_str = "The test case code"
    //Initialise the memory using random data
    for ( p = 0x0; p < MEMORY_SIZE; p += FILE_SIZE ){
        FILE *f = open("file with random data","r");
        mmap(p, f, 0);
    }
    //Initialise the registers using random data
    asm("mov $random, %eax");
    ...
    //Execute the code of the test case
    PC = code_str;
    ((void (*))code_str)();
}
```

Fig. 2. Test procedure example

3.2 Test Case Generation and Execution

Test case is a small piece of assembly programs. We use a hand written xml template to automate the test cases generation procedure, and then define some macros to generate random operands and prefixes of a particular instruction. Both the memory and register are initialized by mapping with random data at first. It's worth noting that we do not need to allocate the entire address space because only part of it may vary. After a preprocessing process, all the macros representing a variable fuzzing component have been replaced with concrete value. We can modify the number of the number of generated test cases directly in our preprocessing procedure, and also customize the test procedure. An example template is in Fig 3.

```
<testcase ring="0">
 <ring0>
  mov $0x200, %eax;
  orl FTG_BITS32, %eax;
  mov %eax, %cr4;
 </ring0>
</testcase>
```

Fig. 3. Template example

We only test a single instruction, and the other instructions are designed for the machine to reach a testing state, which play a role of providing execution contexts for the fuzzed instruction. Our final test cases consist of a bootable kernel, a test case program, some initialization code, which initialize the state of the environment, and transfer the control to the fuzzed instruction. For this reason, we start by executing the test-case program only in one environment and, as soon as the initialization of the state is completed, we replicate the status of the registers and the content of the memory pages to the other environment. Then we execute the code of the test case in the two environments and at the end of the execution, we compare the two final states. Utilizing the built-in snapshot functionality of virtual machine, we can easily get the final state.

The technique we proposed can be used to find VMM crash bugs or the VMM transparency violation defeats. For the latter, we must audit manually to confirm whether it is the VMM's fault or not because it may be a result of undefined behavior in specifications of an ISA.

We choose the 32bit protection mode of CPU to test herein, but for the other modes (the real mode, the virtual 8086 mode, and system mode) which are used less frequently and thus perhaps more buggy resulting from less testing efforts have made, this technique also works with little changes.

4 Conclusion

In this paper, we first introduce the background of VM and VM security, and then figure out the challenges of implementing an IA-32 ISA based VMM. Based on these discussions, we propose a novel test case generation technique for VMM fuzzing, which is focus on testing the critical instructions in IA-32. Benefited from the novel test case generation, the proposal in this paper can find the potential vulnerabilities with fewer test cases compared with other test case generation techniques.

References

1. Ormandy, T.: An Empirical Study into the Security Exposure to Host of Hostile Virtualized Environments. In: Proceedings of CanSecWest Applied Security Conference (2007)
2. Hori, Y., Claycomb, W., Yim, K.: Guest Editorial: Frontiers in Insider Threats and Data Leakage Prevention. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA) 3(1/2) (March 2012)

3. Ho, S.M., Lee, H.: A Thief among Us: The Use of Finite-State Machines to Dissect Insider Threat in Cloud Communications. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA) 3(1/2) (March 2012)
4. England, P., Manferdelli, J.: Virtual machines for enterprise desktop security. Information Security Technical Report 11(4), 193–202 (2006)
5. IBM X-Force Threat Reports, https://www-935.ibm.com/services/us/iss/xforce/trendreports
6. CVE-2010-2070, http://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2010-2070
7. CVE-2009-1244, http://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2009-1244
8. Popek, G., Goldberg, R.: Formal Requirements for Virtualizable 3rd Generation Architectures. Communications of the ACM 17(7), 412–421 (1974)
9. Robin, J.S., Irvine, C.E.: Analysis of the Intel Pentium's Ability to Support a Secure Virtual Machine Monitor. In: Proceedings of the 9th Conference on USENIX Security Symposium (USENIX 2000). USENIX Association, Berkeley (2000)
10. Sutton, M., Greene, A., Amini, P.: Fuzzing: Brute Force Vulnerability Discovery. Addison Wesley Professional (2007)

A Variant of Non-Adaptive Group Testing and Its Application in Pay-Television via Internet*

Thach V. Bui[1], Oanh K. Nguyen[2], Van H. Dang[1], and Nhung T.H. Nguyen[1,3], and Thuc D. Nguyen[1]

[1] University of Science, Ho Chi Minh City, Vietnam
{bvthach,dhvan,nthnhung,ndthuc}@fit.hcmus.edu.vn
[2] Saigon Technology University, Ho Chi Minh City, Vietnam
oanh.nguyenkieu@stu.edu.vn
[3] Japan Advanced Institute of Science and Technology, Japan

Abstract. In non-adaptive group testing (NAGT), the time for decoding is a crucial problem. Given an unknown string $x \in \{0,1\}^N$ with at most d ones, the problem is how to determine $x_i = 1$ using as few tests as possible so that x can be decoded as fast as possible. A NAGT can be represented by a $t \times N$ matrix. Although we do not know x, this matrix, which is called d-disjunct matrix, can reconstruct it exactly. In this paper, we consider a general problem, in which x is an array of N non-negative integer elements and has up to d positive integers. From nonrandom construction, we prove that we can decode a d-disjunct matrix, which is built from $[n,k]_q$-Reed-Solomon codes and identity matrix I_q, and recover x defined above in $poly(d) \cdot t \log^2 t + O(d^3 n \log(d \log N))$ with $t = O(d^2 \log^2 N)$. We also discuss this problem when x contains negative integer elements.

Pay-Television internet-based can be applied these results directly. Since the number of customers is very large, our system must be prevented from illegal buyers. This problem is called *traitor tracing*. To the best of our knowledge, this is the first result that raises a variant of NAGT and gets how to trace traitors without using probability.

Keywords: Group Testing, Traitor Tracing, Pay-TV via Internet.

1 Introduction

In the World War II, the authorities in USA enlisted millions of citizens to join the army. At that time, infectious diseases such as gonorrhea, syphilis, are serious problems. The cost for testing who was infected in turn was very expensive and it also took several times. They wanted to detect who was infected as fast as possible with the lowest cost. R. Dorfman [19], a statistician worked for United States Army Air Forces, proposed that we got N bloods samples from

* This work was financially supported by the KC.01.TN16/11-15, Ministry of Science and Technology (MOST) grant and the National Foundation for Science and Technology Development (NAFOSTED), Vietnam.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 324–330, 2013.

N citizens and combined groups of blood samples to test. It would help him detect infected/disinfected soldiers as few tests as possible. This idea formed a new research field: Group Testing. However, he did not give an explicit way to detect who was infected. D.-Z. Du and F. K. Hwang [18] gave a *naive* algorithm to solve this problem. If the test is negative, all soldiers, whose blood samples belong to this test, are not infected. Otherwise, at least one is infected. When we know who is not infected, the remaining soldiers are infected. For a formal definition, we represent Group Testing as a $t \times N$ binary matrix M, where each column stands for a sample and each row stands for a test. $M_{ij} = 1$ means the jth sample belongs to the ith test, and vice versa. The N infected/disinfected samples are considered as a vector $X = (x_1\ x_2\ ...\ x_N)^T$, where $x_j = 1$ if and only if (iff) jth sample is infected. An outcome vector, or an outcome of testing, is equal to $C = MX$. It is easy to map $C_i \geq 1$ to ith test which is infected. The time to decode C using naive algorithm is $O(tN)$. The decoding time is very important, however, not be considered for the long time. In 2010, P. Indyk, H.Q. Ngo and A. Rudra [1] proved that we could decode d-disjunct matrix in $poly(d) \cdot t \log^2 t + O(t^2)$. They also showed that these d-disjunct matrices were strongly explicit construction, e.g. any entry in M could be computed in time $poly(t)$. The other critical problem in Group Testing is how to generate d-disjunct matrices. There exists two approaches for this problem: probability ($t = O(d^2 \log N)$) and non-randomness ($t = O(d^2 \log^2 N)$). In many high accurate applications, e.g. cryptography, we can not use random construction because we want to control everything. Therefore, nonrandom construction is very important. In this paper, we only consider the nonrandom construction of d-disjunct matrix. For applications, NAGT can be found in data stream [24], data forensics [23] and DNA library screening [25].

List decoding has developed about 50 years. The initial works by [14] defined what list decoding was and gave some bounds for this code. For more details, please refer to the thesis of V. Guruswami [15]. List decoding has many applications and *traitor tracing* is a one of them [16]. Although A. Silverberg et.al. [16] found the relationship between the traitor tracing and list decoding in 2001, traitor tracing had already raised by Chor B. et. al. [12] seven years ago. Traitor tracing is very useful in systems which have pirated users.

Group testing, list decoding and traitor tracing have a strong relationship. In 2010, Indyk P., Ngo H.Q. and Rudra A. [1] proved that list decoding and group testing could be constructed in the same way. Next year, M. Peter, and T. Furon [3] proved that group testing and traitor tracing could be interchangeable since they had the same goal. However, they used probability to solve their problem. Therefore, the output might contain errors.

Digital television (TV) is widely used and studied [8]-[10]. However, when Pay-TV with cable and satellite TV become more and more popular, Pay-TV via internet is also a promising business using the advantages of broad-band networks. There would be a large number of users at the same time for live programs such as football matches or music live shows. One of the threats of this system is their customers can share their account with others. This would lead

to bandwidth overload in rush hours. In 2002, C. Lobbecke et. al. [6] studied German's internet-based TV market. Although the entrance feasibility (include technology) is not clear, they believed that it will be popular. In 2011, according to [7], the world population was 7 billions. Among it, China's and India's population are 1,345.9 and 1,241.3 millions, respectively. Therefore, if only small fraction (assume 0.1%) of their population use Pay-TV internet-based, the number of users is over 1 million. Therefore, the rising problem is how to prevent and detect illegal users in this system. Shuhui HOU, et al. [11] showed that they could detect k colluders (d traitors in our term) with code length k^3 and support up to about k^4 users (N users in our terms). Using group testing, we can treat this problem better than them.

Our Main Result: In this paper, we present a variant of NAGT, show that the traitor tracing problem can be solved without using probability and illustrate these results through an application in Pay-TV via internet. To the best of our knowledge, this is the first result that raises variants of NAGT and gets how to trace traitors without using probability.

Paper Outline: In Section 2, we present some preliminaries and prove the efficient decoding time of the variant of Group Testing. In Section 3, we describe its application in Pay-TV internet-based, compare the efficiency of our proposed solution and raise some open problems. The last Section is conclusion.

2 Preliminaries

2.1 d-disjunct Matrix, Reed-Solomon Codes and Concatenated Codes

An $t \times N$ binary matrix M is a *d-disjunct* matrix iff the union of at most d columns does not contain another columns. The rising problem is how to construct matrix M. Kauzt and Singleton [17] had a strongly explicit way to construct a binary superimposed code of order m based on Unique-Decipherable of order $m+1$ (UD_{m+1}) or Zero-False-Drop of order m (ZFD_m). There has a strong relationship between the binary superimposed code and the disjunct matrix. A matrix M being d-disjunct matrix is equivalent to a binary superimposed code of order d.

G. David Froney Jr. [20] presented a basic knowledge about *concatenated codes*. The concatenated codes are constructed by an *outer code* $C_{out} : [q]^{k_1} \to [q]^{n_1}$, where $q = 2^{k_2}$, and a binary *inner code* $C_{in} : \{0,1\}^{k_2} \to \{0,1\}^{n_2}$. Let $C = C_{out} \circ C_{in}$ be a concatenated code. C's size is $(n_1 n_2) \times q^{k_1}$. In [17], the authors chose C_{out} as a q-nary code and C_{in} as an identity matrix.

Reed-Solomon (RS) codes, which were invented by Reed, I.S. and Solomon, G. [21], are the famous codes that are applied in many fields [22]. They are not only q-nary codes but also the *maximum distance separable* codes. A $[n,k]_q$-code C, $1 \le k \le n \le q$, is a subset $C \subseteq [q]^n$ of size q^k. The parameters n, k and q are known as the *block length, dimension* and *alphabet size*. In this model,

we choose C_{out} as $[q-1,k]_q$-RS code and C_{in} as an identity matrix I_q. A d-disjunct matrix ($d = \lfloor \frac{n-1}{k-1} \rfloor$) is achieved from $C_{out} \circ C_{in}$ by putting all $N = q^k$ codewords as columns of the matrix. According to [17], given d and N, if we chose $q = O(d \log N)$, $k = O(\log N)$, the resulting matrix is $t \times N$ d-disjunct, where $t = O(d^2 \log^2 N)$. In 2010, P. Indyk, Hung Q. Ngo and A. Rudra [1] gave a random construction of d-disjunct matrices with $t = O(d^2 \log N)$ and cited to [17] for non-random construction with $t = O(d^2 \log^2 N)$, that can be decoded in $poly(d) \cdot t \log^2 t + O(t^2)$.

2.2 A Variant of Group Testing and the Connection between Traitor Tracing and Group Testing

In 2011, P. Meerwald and T. Furon [3] pointed out that there exists a connection between Group Testing and Traitor Tracing. Researchers in these fields aim to find very few specific people in a huge population. The authors used probability to illustrate their model. After estimating d (K in [3]), they computed $Score_j$ for jth user and checked whether $Score_j$ was larger than a threshold. If the answer is "Yes", jth user is infected, and vice versa. This procedure is called *single decoder with likelihood ratio test*. Since this method was relied on probability and the threshold, which was "ambiguous" parameter, their result might contain some false positive users. To improve the performance, the authors calculated scores for τ-tuples and reduced significantly the number of tests in practice. However, in general, since the probability of exact recovery is never equals to 1, using this solution may lead us accuse wrong users. Moreover, the time to find infected users is still questionaire for this approach.

The main question in Group Testing is how to identify the positive values in **binary** vector x from $y = Mx$, where $M_{t \times N}$ is d-disjunct and $x = (x_1\ x_2 \cdots x_N)^T$. The value x_i represents ith person. If this individual is infected iff $x_i = 1$. In 2012, T.D. Nguyen, T.V. Bui, V.H. Dang and D. Choi [2] extended this problem turn out: given $M_{t \times N}$ is d-disjunct and $y = Mx$, where $x = (x_1\ x_2 \cdots x_N)^T$ is a **non-negative** vector and $wt(x) = \sum_{i=1, x_i \neq 0}^{N} 1 \leq d$, find x from y and M.

For the convenience, the phrase "the frequency of jth user" and "the frequency of his keys" are equivalent. We also denote I_j is the jth column of I_q and $M_x(y)$ is (x, y) entry of $[n,k]_q$-RS codes. The following theorem describes the efficient decoding time of variant of Group Testing.

Theorem 1. *If any d-disjunct matrix $M_{t \times N}$ is constructed by concatenated codes, which is built from $[n,k]_q$-RS codes and identity matrix I_q, we can recover a non-negative integer vector $x_{N \times 1}$ from $y = Mx$ in $poly(d) \cdot t \log^2 t + O(d^3 n \log(d \log N))$, where $wt(x) \leq d$.*

Proof. According to Corollary C.1 [1], $[n,k]_q$-RS codes are $(d, O(d^2 \log(d \log N)))$-list recoverable codes ($N = q^k$ and $q = d \log N$). They can be decoded in $poly(d,q)$ [5] or $poly(d).t \log^2 t$ [4], and output $\mu = O(d^2 \log(d \log N))$ candidates. We denote the index of μ candidates of x is $\tau = \{\tau_1, \tau_2, \cdots, \tau_t\}$. After

that, we split the result vector y into n blocks $C = \{S_1, S_2, \ldots S_n\}$, each block's size is q. For each block, we decompose this block into set of symbols (in F_q) by the following rule: If $S_i = f_{i_1} I_{i_1} + f_{i_2} I_{i_2} + \ldots + f_{i_l} I_{i_l}$, where $f_{i_k} > 0$ for $k = 1, 2, \ldots, l$, then S_i can be represented as follow: $S_i = \{\{i_1, f_{i_1}\}, \{i_2, f_{i_2}\}, \ldots, \{i_l, f_{i_l}\}\}$, where $i = 1, 2, \cdots, n$. According to [2], the frequency of τ_jth user is $\min\{f_{i_k} : M_i(\tau_j) = i_k \in S_i,\ \forall i = 1, \ldots, n\}$. τ_jth user is infected iff $f_{\tau_j} \neq 0$. Since $|S_i| \leq d$, the time for finding infected users is $dn.O(d^2 \log(d \log N))$. Therefore, the overall time is: $poly(d) \cdot t \log^2 t + O(d^3 n \log (d \log N))$.

3 The Application in Pay-TV Internet-Based and Comparison

In [12], Chor B. et. al. proposed the scheme that although server only keeps t keys, it could support N users, where $t \ll N$. The key idea is each user holds *a set of keys* F, $|F| \leq t$. A pirated user will use a set of keys, that may be combined from a small group of traitors. The authors trace the traitors by using probability, which may create error tolerant. To overcome this issue, we propose a method that can find exactly who are traitors and how many times their keys are used.

3.1 Algorithm

In this scenario, at the time we check whether who is traitor, assume that all users log in our system and the pirated users are at most d. A d-disjunct matrix $M_{t \times N}$ is generated by the system. Assume $sum = M \times 1_{N \times 1}$ and $C = 0_{t \times 1}$. Every jth user is represented by a unique column M_j of M. Server stores t keys, denote that $F_{key} = \{k_1, k_2, \cdots, k_t\}$ and jth user stores a subset $F_j = \{k_h : M_{hj} = 1,\ where\ h = 1, \ldots, t\}$ of F_{key}. The key distribution procedure can apply the way that D. Boneh and M.Franklin proposed in [13]. When jth user is authenticated, server will increase the counter $C = C + M_j$. After this procedure, we calculate $trace = C - sum$ and use the Theorem 1 to find who is traitor and the frequency of his keys.

If jth user is disinfected, the frequency of the set F_j is 1. Therefore, if all users are disinfected, the vector counter will be equal to $C = \sum_{j=1}^{N} M_j = sum$. Note that there exists the bijection between M_j and F_j, if any hth user is infected, C will turn out: $C = sum + M_h$. Thus, after authentication phrase, if $trace = C - sum$ is not equal to zero, some users are counterfeited.

3.2 Comparison

In practice, the authors [11] only generate up to $p = 11$ (d in our term), the code length is 1332 and support at most 13431 users. Using MATLAB, we can generate d-disjunct matrices as defined in Section 2 and support the number of users as much as we want. For examples, a matrix that is generated from $[31, 3]_{32}$-RS codes and identity matrix I_{32} can support up to $32^3 = 32768$ users, detect at most $d = \lfloor \frac{31-1}{3-1} \rfloor = 15$ where the code length is $t = 31 \times 32 = 992$.

In theory, since the authors built ACC code from BIBD code, they faced many problems from this approach. D.-Z. Du and F. K. Hwang [18] pointed out that for the same d and N, the code length that was achieved from random construction is always smaller than BIBD construction. Last, in [11], the authors did not show how to find p colluders. In our solution, we satisfy this requirement as well. In the above model, we are only successful if the number of authenticated users is larger than N and all users who are legal must be log in your system. It seems to be practical since users pay money for this service. However, though the number of users is larger than N, some legal users are missing (they do not log in at that time). Hence, there are some illegal users. In this case, how can we identify them when the system that contains both pirated and missing users? The following proposition will answer who traitors are and also be *other variant* of NAGT.

Proposition 2. *Given d-disjunct matrix $M_{t\times N}$ is constructed by concatenated codes, which is built from $[n,k]_q$-RS code and identity matrix I_q, the positive integer vector $x_{N\times 1} = (x_1, x_2, \ldots, x_N)^T$ and binary vector $y_{N\times 1}$ such that: $x_i > wt(y)$ if $x_i > 0$ for $i \in [N]$ and $wt(x) + wt(y) \leq d$. If $z = M(x - y)$, we can identify the index of positive elements of x in time poly(t).*

Proof. Let denote $\gamma = x - y$. Since $wt(x) + wt(y) \leq d$, $|\gamma| \leq d$. Assume the index set of positive elements of x is $I = \{i_1, i_2, \ldots, i_h\}$. Thanks to $x_k > wt(y)$ where $k \in I$, the positive elements of x are also the positive elements of z. The output vector $z = (z_1, z_2, \ldots, z_t)^T$ will be converted in to positive/negative vector by the following rule: ith test is positive iff $z_i > 0$. After this conversion, using the Corollary C.1 in [1] to find the index of positive elements of x.

4 Conclusion

We present for the first time the variant of NAGT and the connection between traitor tracing and group testing without using probability. Simultaneously, we show that these results can be applied in Pay-TV via internet. Our future work aims at lowering the cost of decoding d-disjunct matrices which are constructed randomly, detect who is missing and find the frequency of pirated users in Pay-TV internet-based.

References

1. Indyk, P., Ngo, H.Q., Rudra, A.: Efficiently decodable non-adaptive group testing. In: Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1126–1142 (2010)
2. Nguyen, T.D., Bui, T.V., Dang, V.H., Choi, D.: Efficiently Preserving Data Privacy Range Queries in Two-Tiered Wireless Sensor Networks. In: 2012 9th International Conference on Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC), pp. 973–978. IEEE (2012)
3. Meerwald, P., Furon, T.: Group testing meets traitor tracing. In: 2011 IEEE Inter. Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4204–4207 (2011)

4. Alekhnovich, M.: Linear Diophantine equations over polynomials and soft decoding of Reed-Solomon codes. In: Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science, pp. 439–448. IEEE (2002)
5. Parvaresh, F., Vardy, A.: Correcting errors beyond the Guruswami-Sudan radius in polynomial time. In: 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2005). IEEE (2005)
6. Lbbecke, C., Falkenberg, M.: A framework for assessing market entry opportunities for internet-based TV. Inter. J. on Media Management 4(2), 95–104 (2002)
7. Population Reference Bureau: Population data sheet. Population Reference Bureau, Washington, DC (2011)
8. Huang, Y.-L., Shieh, S., Ho, F.-S., Wang, J.-C.: Efficient key distribution schemes for secure media delivery in pay-TV systems. IEEE Transactions on Multimedia 6(5), 760–769 (2004)
9. Macq, B.M., Quisquater, J.-J.: Cryptology for digital TV broadcasting. Proceedings of the IEEE 83(6), 944–957 (1995)
10. Kim, C., Hwang, Y., Lee, P.: Practical pay-TV scheme using traitor tracing scheme for multiple channels. Information Security Applications, 264–277 (2005)
11. Hou, S., Uehara, T., Satoh, T., Morimura, Y., Minoh, M.: Fingerprinting codes for Internet-based live pay-TV system using balanced incomplete block designs. IEICE Transactions on Information and Systems 92(5), 876–887 (2009)
12. Chor, B., Fiat, A., Naor, M.: Tracing Traitors. In: Desmedt, Y.G. (ed.) CRYPTO 1994. LNCS, vol. 839, pp. 257–270. Springer, Heidelberg (1994)
13. Boneh, D., Franklin, M.K.: An Efficient Public Key Traitor Scheme (Extended Abstract). In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 338–353. Springer, Heidelberg (1999)
14. Elias, P.: List decoding for noisy channels. Massachusetts Institute of Technology. Research Laboratory of Electronics (1957)
15. Guruswami, V.: List Decoding of Error-Correcting Codes. LNCS, vol. 3282. Springer, Heidelberg (2004)
16. Silverberg, A., Staddon, J., Walker, J.L.: Efficient Traitor Tracing Algorithms Using List Decoding. In: Boyd, C. (ed.) ASIACRYPT 2001. LNCS, vol. 2248, pp. 175–192. Springer, Heidelberg (2001)
17. Kautz, W.H., Singleton, R.C.: Nonrandom binary Superimposed codes. IEEE Transactions on Information Theory 10(4), 363–377 (1964)
18. Dingzhu, D., Hwang, F.: Combinatorial group testing and its applications. World Scientific Publishing Company Incorporated (1993)
19. Dorfman, R.: The detection of defective members of large populations. The Annals of Mathematical Statistics 14(4), 436–440 (1943)
20. Froney Jr., G.D.: Concatenated codes. DTIC Document (1965)
21. Reed, I.S., Solomon, G.: Polynomial codes over certain finite fields. Journal of the Society for Industrial and Applied Mathematics 8(2), 300–304 (1960)
22. Wicker, S.B., Bhargava, V.K.: Reed-Solomon codes and their applications. Wiley-IEEE Press (1999)
23. Goodrich, M.T., Atallah, M.J., Tamassia, R.: Indexing Information for Data Forensics. In: Ioannidis, J., Keromytis, A.D., Yung, M. (eds.) ACNS 2005. LNCS, vol. 3531, pp. 206–221. Springer, Heidelberg (2005)
24. Cormode, G., Muthukrishnan, S.: What's hot and what's not: tracking most frequent items dynamically. In: Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 296–306. ACM (2003)
25. Ngo, H.Q., Du, D.Z.: A survey on combinatorial group testing algorithms with applications to DNA library screening. Discrete Mathematical Problems with Medical Applications 55, 171–182 (2000)

A Proposal on Security Case Based on Common Criteria

Shuichiro Yamamoto[1], Tomoko Kaneko[2], and Hidehiko Tanaka[3]

[1] Nagoya University
syamamoto@acm.org
[2] NTT DATA CORPORATION
knktmk204th@gmail.com
[3] Institute of Information Security
tanaka@iisec.ac.jp

Abstract. It is important to assure the security of systems in the course of development. However, lack of requirements analysis method to integrate security functional requirements analysis and validation in upper process often gives a crucial influence to the system dependability. For security requirements, even if extraction of menaces was completely carried out, insufficient countermeasures do not satisfy the security requirements of customers.

In this paper, we propose a method to describe security cases based on the security structures and threat analysis. The security structure of the method is decomposed by the Common Criteria (ISO/IEC15408).

Keywords: Security Case, Security Requirements Analysis, Common Criteria.

1 Introduction

It is important to show how a request such as "The system is acceptably secure" is supported by objective evidence for customers. We show the description method by using Assurance Case and Common Criteria as the objective evidence.

In Chapter 2 "Related work," we explain assurance case [1-4] and security case approaches [6-8], as well as an overview of common criteria (CC) [5]. In Chapter 3, we show security case reference patterns based on CC. In Chapter 4, some considerations on the method are described. Chapter 5 explains future issues.

2 Related Work

2.1 Assurance Case

Security case is an application of Assurance case, which is defined in ISO/IEC15026 part 2. Security cases are used to assure the critical security levels for target systems. Standards are proposed by ISO/IEC15026 [2] and OMG's Argument Metamodel (ARM) and [3] Software Assurance Evidence Metamodel (SAEM) [4]. ISO/IEC 15026 specifies scopes, adaptability, application, assurance case's structure and contents, and deliverables. Minimum requirements for assurance case's structure and

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 331–336, 2013.
IFIP International Federation for Information Processing 2013

contents are: to describe claims of system and product properties, systematic argumentations of the claims, evidence and explicit assumptions of the argumentations; to structurally associate evidence and assumptions with the highest-level claims by introducing supplementary claims in the middle of a discussion. One common notation is Goal Structuring Notation (GSN) [1], which widely used in Europe for about ten years to verify system security and validity after identifying security requirements.

2.2 Security Case

Goodenough, Lipson and others proposed a method to create Security Assurance case [6]. They described that the Common Criteria provides catalogs of standard Security Functional Requirements and Security Assurance Requirements. They decomposed Security case by focusing on the process, such as requirements, design, coding, and operation. The approach did not use the Security Target structure of the CC to describe Security case.

Alexander, Hawkins and Kelly overviewed the state of the art on the Security Assurance cases [7]. They showed the practical aspects and benefits to describe Security case in relation to security target documents. However they did not provide any patterns to describe Security case using CC.

Kaneko, Yamamoto and Tanaka recently proposed a security countermeasure decision method using Assurance case and CC [8]. Their method is based on a goal oriented security requirements analysis [9-10]. Although the method showed a way to describe security case, it did not provide Security case graphical notations and the seamless relationship between security structure and security functional requirements.

2.3 Common Criteria

Common Criteria (CC: equivalent to ISO/IEC15408) [5] specifies a framework for evaluating reliability of the security assurance level defined by a system developer. In Japan, the Japan Information Technology Security Evaluation and Certification Scheme (JISEC) is implemented to evaluate and authenticate IT products (software and hardware) and information systems. In addition, based on CC Recognition Arrangement (CCRA), which recognizes certifications granted by other countries' evaluation and authorization schemes, CC accredited products are recognized and distributed internationally. As an international standard, CC is used to evaluate reliability of security requirements of functions built using IT components (including security functions). CC establishes a precise model of Target of Evaluation (TOE) and the operation environment. And based on the security concept and relationship of assets, threats, and objectives, CC defines ST (Security Target) as a framework for evaluating TOE's Security Functional Requirement (SFR) and Security Assurance Requirement (SAR). ST is a document that accurately and properly defines security functions implemented in the target system and prescribes targets of security assurance. ST is required for security evaluation and shows levels of adequacy in TOE's security functions and security assurance.

3 Security Case Reference Patterns

3.1 Issues to Describe Security Case

Product and process are both important to assure system security. In this paper we propose a hierarchical method to describe Security case. We decompose Security case based on Security Target structure in the upper part. And then we describe bottom part of the Security case based on security analysis process.

3.2 Security Case Based on Security Target Structure

Fig.1. describes an example pattern for Security case based on CC.

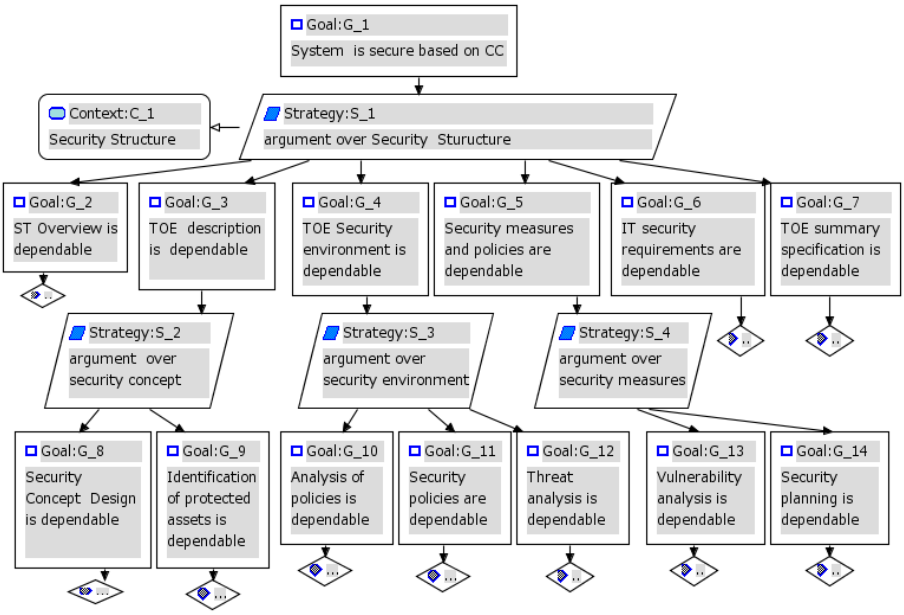


Fig. 1. Security case pattern for CC based Security Structure

The figure shows the Security Target structure and security analysis process consists of the two decomposition layers. In the first decomposition, ST overview, TOE description, TOE security environment, Security measures and policies, IT security requirements, and TOE summary specification are described. For each decomposed claim, arguments are also attached to decompose it by security analysis process. For example, to assure the dependability of the TOE security environment, the security analysis process is decomposed by three claims, i.e., Analyzing protection policies, Clarifying security policies, and threat analysis.

3.3 Security Case to Assure Security Requirements against Threats

Fig.2. describes security case to assure security functional requirements. It consists of the following hierarchical layers, Threats category, Activity of threats, and Security function layers. The security case can be considered as the decomposition of the claim G_6 in Fig.1.

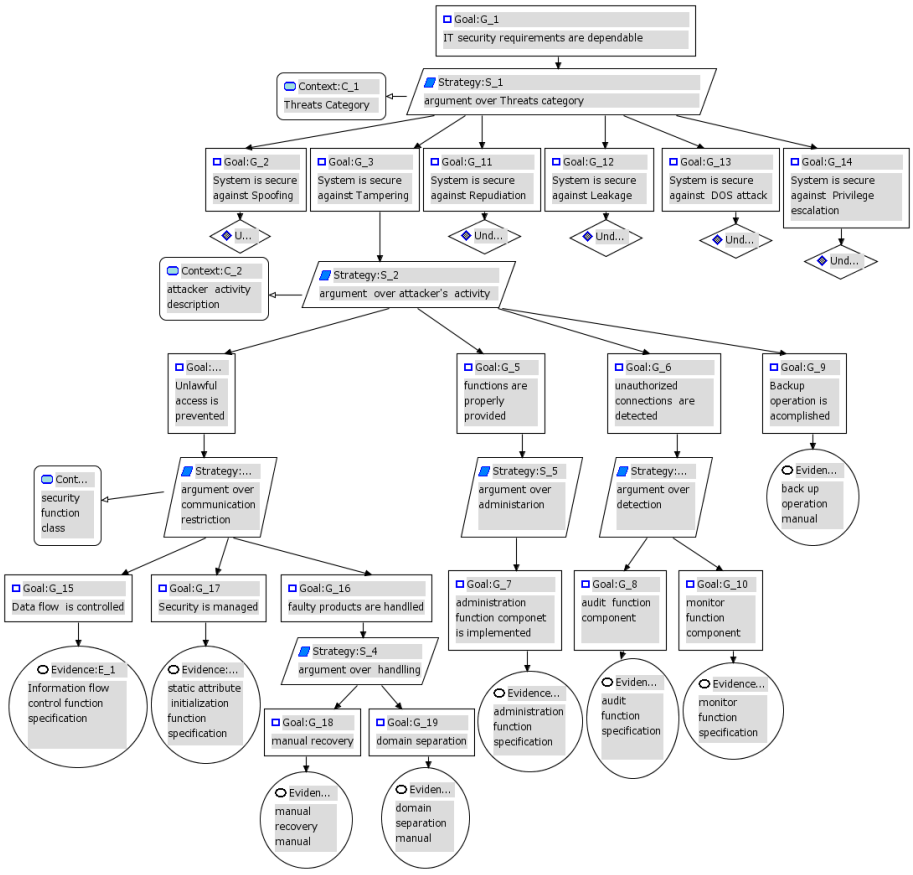


Fig. 2. Security case pattern for security function specification based on CC

The sample case is created based on PP [13] provided by IPA (Information-technology Promotion Agency) and is not the example of actual specific system. Thus, the sample case should be regarded as a reference model of Security case.

4 Considerations

Describing Security case according to ST structure of CC has an advantage in validating objective assurance levels based on an international standard notation. It is

possible to properly define and implement security functions in line with ST structure and appropriate threat analysis. We also can implement negotiated security functions based on structured way of Security case and international standardized terminologies in CC of catalogued security function levels.

The relationship between security structure of CC and Security case structure is mandatory for compatibility. As shown in the examples of section 3, the Security case structure is seamlessly correspondent to CC.

We also confirmed a way to integrate Security cases between Security Target structure and Security functional requirements as shown in the goal relationship of two figures.

5 Future Issues

There are some unsolved issues in security case development presented in this paper.

Our study is still in a preliminary phase and further evaluation needs to be done in future. It is necessary to evaluate the proposed method for designing actual system development. The proposed approach provides a reference Security case structure. Therefore, it can also be used to effective validation of the target systems compatibility to CC. This kind of application of our method will provide a simple integration process between security design and validation.

We also have a plan to develop Security case patterns based on this paper. This will ease to reuse Security cases based on CC. This research is an extension of Safety case pattern proposed by Kelly and McDermid [11].

In terms of CC based security requirement analysis, goal oriented methods and use-case based methods are proposed [12]. Therefore, it is desirable to verify effectiveness of our method by comparing our method with these methods.

References

1. Kelly, T., Weaver, R.: The Goal Structuring Notation – A Safety Argument Notation. In: Proceedings of the Dependable Systems and Networks 2004 Workshop on Assurance Cases (2004)
2. ISO/IEC15026-2-2011,Systems and Software engineering-Part2: Assurance case
3. OMG, ARM, http://www.omg.org/spec/ARM/1.0/Beta1/
4. OMG, SAEM, http://www.omg.org/spec/SAEM/1.0/Beta1/
5. Common Criteria for Information Technology Security Evaluation, http://www.commoncriteriaportal.org/cc/
6. Goodenough, J., Lipson, H., Weinstock, C.: Arguing Security - Creating Security Assurance Cases (2007), https://buildsecurityin.us-cert.gov/bsi/articles/knowledge/assurance/643-BSI.html
7. Alexander, T., Hawkins, R., Kelly, T.: Security Assurance Cases: Motivation and the State of the Art, CESG/TR/2011 (2011)
8. Kaneko, T., Yamamoto, S., Tanaka, H., Proposal on Countermeasure Decision Method Using Assurance Case And Common Criteria. In: ProMAC 2012 (2012)

9. Kaneko, T., Yamamoto, S., Tanaka, H.: SARM – a spiral review method for security requirements based on Actor Relationship Matrix. In: ProMAC 2010, 1227–1238 (2010)
10. Kaneko, T., Yamamoto, S., Tanaka, H.: Specification of Whole Steps for the Security Requirements Analysis Method (SARM)- From Requirement Analysis to Countermeasure Decision. In: ProMAC 2011 (2011)
11. Kelly, T., McDermid, J.A.: Safety Case Construction and Reuse using Patterns. In: Proceedings of 16th International Conference on Computer Safety, Reliability and Security. In: SAFECOMP 1997. Springer (September 1997)
12. Saeki, M., Kaiya, H.: Security Requirements Elicitation Using Method Weaving and Common Criteria. In: Chaudron, M.R.V. (ed.) MODELS 2008. LNCS, vol. 5421, pp. 185–196. Springer, Heidelberg (2009)
13. http://www.ipa.go.jp/security/fy13/evalu/pp_st/pp_st.html

An Adaptive Low-Overhead Mechanism for Dependable General-Purpose Many-Core Processors

Wentao Jia*, Rui Li, and Chunyan Zhang

School of Computer, National University of Defense Technology, China
{wtjia,lirui,cyzhang}@nudt.edu.cn

Abstract. Future many-core processors may contain more than 1000 cores on single die. However, continued scaling of silicon fabrication technology exposes chip orders of such magnitude to a higher vulnerability to errors. A low-overhead and adaptive fault-tolerance mechanism is desired for general-purpose many-core processors. We propose high-level adaptive redundancy (HLAR), which possesses several unique properties. First, the technique employs selective redundancy based application assistance and dynamically cores schedule. Second, the method requires minimal overhead when the mechanism is disabled. Third, it expands the local memory within the replication sphere, which heightens the replication level and simplifies the redundancy mechanism. Finally, it decreases bandwidth through various compression methods, thus effectively balancing reliability, performance, and power. Experimental results show a remarkably low overhead while covering 99.999% errors with only 0.25% more networks-on-chip traffic.

Keywords: Many-Core, Redundant Execution, Adaptive Dependable, Low-Overhead.

1 Introduction

Transistors continue to double in number every two years without significant frequency enhancements and extra power costs. These facts indicate a demand for new processors with more than 1000 cores and an increasing need to utilize such a large amount of resources [1]. As transistor size decreases, the probability of chip-level soft errors and physical flaws induced by voltage fluctuation, cosmic rays, thermal changes, or variability in manufacturing further increases [2], which causes unavoidable errors in many-core systems.

Redundant execution efficiently improve reliability, which can be applied in most implementations of multithreading such as simultaneous multithreading (SMT) or chip multi-core processors (CMP). Current redundant execution techniques such as SRT, CRT and RECVF [3] entail either high hardware costs such

* This work was Supported by the National Nature Science Foundation of China under NSFC No. 61033008, 60903041 and 61103080.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 337–342, 2013.

as load value queue (LVQ), buffer, comparer, and dedicated bus, or significant changes to existing highly optimized micro-architectures, which may be affordable for CMP but not for general-purpose many-core processors (GPMCP). Same core-level costs in many-core processors result in an overhead up to 10 or 100 times higher than that in CMP. High overhead may be reasonable for fixed high-reliable applications but not for all applications in GPMCP as the latter induce large overhead for applications that do not require reliability.

GPMCP present two implications. First, not all applications require high reliability. Second, the chip contains more than 1000 cores and some cores usually become idle. We proposed high-level adaptive redundancy (HLAR), an adaptive, low-overhead mechanism based application assistance and system resource usage for GPMCP. Our evaluation-based error injection shows that HLAR is capable of satisfying error coverage with minimal NoC traffic that covers 99.999% errors with 0.25% more NoC traffic.

2 Background and Related Work

Many-core architectures such as Tilera Tile64, Intel MIC, and NVIDIA Fermi, among others show good perspective. A challenge for many-core architecture is that hardware-managed Cache entail unacceptable costs. As alternative model, software-managed local memory (LM), exposes intermediate memories to the software and relies on it to orchestrate the memory operations. The IBM C64 is a clear example of a device that benefits from the use of LMs.

Applications in GPMCP are parallel and consist of threads with different reliability requirements. Some applications are critical. Wells [4] managed two types of applications: reliable applications and performance applications. Kruijf [5] argued that some emerging applications are error-tolerant and can discard computations in the event of an error.

The following have prompted this study: DCC [6] allows arbitrary CMP cores to verify execution of the other without a static core binding or a dedicated communication hardware. Relax [5] proposed a software recovery framework in which the code regions are marked for recovery. Fingerprinting [7] proposed compressing architectural state updates into a signature, which lowers comparison bandwidth by orders of magnitude.

3 HLAR Mechanism

Redundancy overhead. HLAR redundancy are not executed for the whole chip, thus, we define *redundancy entry* as cores that execute redundancy and *outside redundancy entry* as cores that do not execute redundancy. Considering processors without redundancy as the baseline, we classify redundancy overhead.

(i) Fixed overhead in all cores because of hardware cost or performance lost due to hardware modification (O_{FA}).

(ii) Special overhead in redundancy entry (O_{SRE}).

(iii) Temporal overhead in redundancy entry (O_{TRE}).

(iv) Overhead outside redundancy entry due to bandwidth and other shared resources (O_{ORE}). If redundancy cores utilize additional shared resources, other cores may cost more to obtain the resources. NoC bandwidth is a global resource and affects the performance of the die.

HLAR Architecture. HLAR is a low overhead GPMCP redundancy mechanism It supports redundancy between arbitrary cores and causes minimal modifications to the core. Fig.1 shows the HLAR architecture. HLAR employs a many-core processor interconnected with a mesh network. For each core, we added a redundancy control module called redundancy manager (RM). RM consists of small control circuits and two FIFO buffers. The RM area is small, compared with cores/networks. HLAR uses the existing chip networks in transferring trace data and supports input replication via a remote value queue (RVQ).

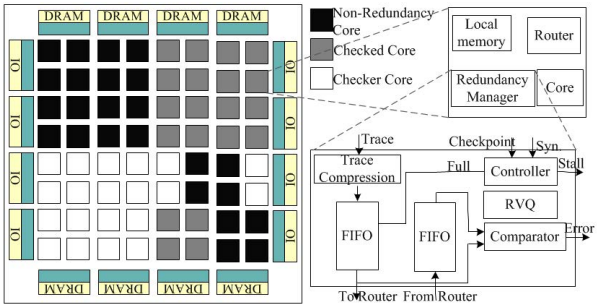


Fig. 1. HLAR architecture

HLAR cores can be a checked core, a checker core, or a non-redundancy core. Non-redundancy cores do not execute redundancy. In the checked core, the RM receives register updates from the core and writes these updates onto the sender FIFO. Typically, the register update information consists of address/value pairs (8bit/32bit). The RM interfaces with the NoC, compresses the trace data, and sends messages to the checker core. The RM of the checker core writes its compressed register updates into its sender FIFO. A comparator compares elements in the two FIFOs. When these two vary, the RM raises a redundancy error exception. The RM controller includes a simple state-machine that stalls the processor when the FIFOs become full.

Input Replication. Unlike the previous work, only the remote value but the load value requires replication in HLAR. HLAR supports input replication via RVQ. The checker core reads values from the RVQ rather than access the remote node. Programs and data in the checked core's local memory must be copied onto the checker core's, thus, address space replication is needed. A register named *reg_addr_remap* is used to replicate the address space.

Trace Compression. We employed CRC-1, CRC-5, and CRC-8 to obtain an adaptive compression rate, and found that these methods adequately satisfy

coverage. To obtain more effective compression, HLAR summarizes many updates for once CRC check. CRC-1/10 means one-bit CRC-1 check for every set of 10 trace values, hence, CRC-5/100, CRC-8/1000 and so on. NoC overhead is very small in HLAR(for example, CRC-8/1000 only costs 0.025% more bandwidth)

Recovery. Redundant execution is often combined with a checkpoint retry to gain recovery. However, the checkpoint overhead increases, especially if a short checkpoint interval is employed. HLAR indicates the *checkpoint_interval* for a configurable checkpoint mechanism. Aside from the checkpoint, a simple forward error recovery (FER) mechanism is employed, which discards incorrect results and continues to make progress.

Application Framework in HLAR. HLAR for applications can be as simple as system devices, in which only require configuring and enabling. The device views simple usage in applications and management in system. The application first configures HLAR through HLAR_config() and the control registers in RM are then set. When HLAR_enable() is prompted, the Hypervisor selects the appropriate core, copies the state from the checked core to initialize the checker core, and then begins the redundant execution. The hypervisor completes the redundancy and disables the RM until HLAR_disable() is called.

4 Evaluation

4.1 Methodology

HLAR is implemented based on the OpenRISC 1200 (OR1200) [8] core. The OR1200 configuration is used as the default parameter, except for the following: 8 KB local instructions Cache, 8 KB local data Cache, and 32 KB local memory.

Our experimental results are based on OR1200 cores in Altera Stratix IV FPGA prototyping boards. The primary advantage of using prototyping boards instead of simulation is speed. We evaluated the error coverage and latency by error injection and evaluated the overhead based hardware counters and EDA tools(Quartus II 11.0).

Workload and Fault Model. The applications evaluated include the following: MatrixMult (4M) and BubbleSort (5M). The fault model consists of single bit-flip (95%) and double bit-flip (5%). The number of experiments is 20,000 + 40,000 + 100,000 (fault sites for CRC-1, CRC-5, and CRC-8) * 6 (summarized interval) * 2 (applications) = 1,920,000.

Error Injection. One register from the 32 GPRs and 6 control registers (PC, SR, EA, PICMR, PICPR, and PICSR) were randomly chosen. Likewise, 1- or 2-bit out of 32-bit locations were randomly selected and flipped.

4.2 Experimental Results

Temporal Overhead in Redundancy Entry. O_{TRE} is usually shown as performance degradation. Performance degradation shown in Fig.2(a) is 0.71% for

MM and 0.68% for BS at 100,000 instructions of the checkpoint. These rates increase to 12.2% and 9.8% at 1000 instructions. When a discard recovery mechanism is employed, the degradation is negligible at 0.42% and 0.23%, as shown in Fig.2(b).

Fixed Overhead. The only fixed overhead in HLAR is RM. Logic utilization is shown in Fig.2(c). RM utilizes 359 combinational ALUTs and a 160 byte memory, which only use 2.46% and 0.25% of the total (OR1200 core and router). The fixed overhead in HLAR is much less compared with Reunion or RECVF.

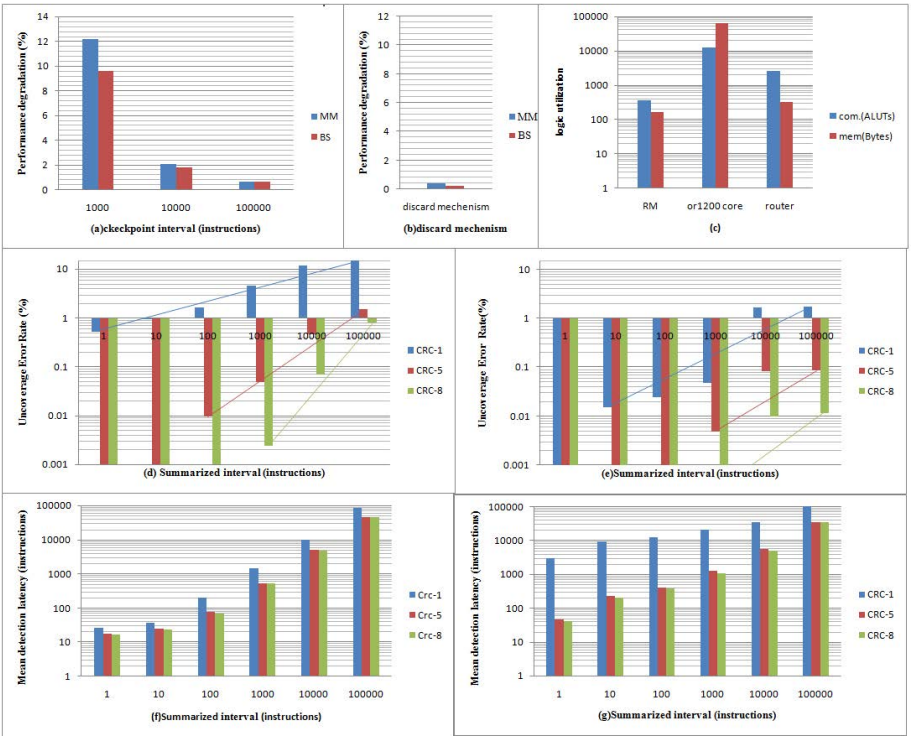


Fig. 2. Result: (a)Temporal overhead for checkpoint and for (b)discard mechanism; (c)Fixed overhead; (d)Error coverage rate for MM and (e)BS; (f)Mean detection latency for MM and (g)BS.

Error Coverage. HLAR compresses traces with CRC-1, CRC-5, CRC-8, and summarizes CRC to balance reliability and NoC performance. The uncoverage error rate is shown in Fig.2(d) and (e). Comparing traces without compression can obtain a 100% coverage. CRC-1/1 reduces bandwidth by up to 40 times without losing coverage, 0.53% for MM, and below 0.001% for BS. When the summarized interval increases by 10 times, uncoverage increases (denoted by a line) on a small scale. Uncoverages in CRC-5 and CRC-8 are low even at intervals of 10,0000 instructions; uncoverage rates are 1.5% and 0.8%, respectively,

for MM and are 0.086% and 0.012% for BS. As the interval decreases, uncoverage decreases significantly. An uncoverage rate below 0.001% in Fig.2(d) and (e) indicates that no SDC occurs after error injection. Minimal uncoverage (below 0.001%) occurs at 10 and 100 instructions in CRC-5 and at 100 and 1000 instructions in CRC-8 in MM and BS, respectively, which means only 0.25% or 0.025% more NoC traffic.

Detection Latency. NoC communicating with a distant core may incur greater latency than communicating with an adjacent core. The summarized CRC may also lead to larger latency. However, the results in Fig.2(f) and (g) show that the detected latency in HLAR is bounded. Mean error detection latency (MEDL) for MM is consistent with the summarized interval, increasing from 27 in CRC-1/1 to 86,490 in CRC-1/100000. CRC-5 and CRC-8 show lower MEDL than CRC-1. For instance, at the interval of 1000 instructions, MEDL is 1424 in CRC-1, 539 in CRC-5, and 515 in CRC-8.

5 Conclusion

We analysed the redundant execution overhead and proposed HLAR, an adaptive low-overhead redundancy mechanism for GPMCP. Unlike prior mechanisms, HLAR can sense application requirements and system resource usage to reconfigure redundancy. Thus, HLAR decreases the overhead by executing only the necessary redundancy and using the idle core for this redundancy. HLAR expands the local memory within the replication sphere, which provides relaxed input replication, distributes the memory access, and allows the core pairs to progress simultaneously. HLAR is capable of perfect error coverage with a minimal overhead, covering 99.999% of errors with less than 0.25% more commutation.

References

1. Borkar, S.: Thousand core chips: a technology perspective. In: Proceedings of the 44th Annual Design Automation Conference (June 2007)
2. Srinivasan, J., Adve, S.V., Bose, P., Rivers, J.A.: The impact of technology scaling on lifetime reliability. In: Intl. Conf. on DSN (June 2004)
3. Subramanyan, P., Singh, V., Saluja, K.K., Larsson, E.: Energy-Efficient Fault Tolerance in Chip Multiprocessors Using Critical Value Forwarding. In: Intl. Conf. on Dependable Systems and Networks (June 2010)
4. Wells, P.M., Chakraborty, K., Sohi, G.S.: Mixed-mode multicore reliability. In: Intl. Conf. on ASPLOS (March 2009)
5. de Kruijf, M., Nomura, S., Sankaralingam, K.: Relax: An architectural framework for software recovery of hardware faults. In: ISCA (2010)
6. LaFrieda, C., Ipek, E., Martinez, J.F., Manohar, R.: Utilizing dynamically coupled cores to form a resilient chip multiprocessor. In: Intl. Conf. on DSN (2007)
7. Smolens, J.C., Gold, B.T., Kim, J., Falsafi, B., Hoe, J.C., Nowatzyk, A.G.: Fingerprinting: bounding soft-error detection latency and bandwidth. In: Intl. Conf. on ASPLOS (October 2004)
8. Lampret, D.: OpenRISC 1200 IP Core Specification (September 2001), `http://www.opencores.org`

Identity Management Lifecycle - Exemplifying the Need for Holistic Identity Assurance Frameworks

Jostein Jensen

Norwegian University of Science and Technology, Department of Computer and Information Science, Norway
jostein.jensen@idi.ntnu.no

Abstract. Many governments around the world have a strategy to make electronic communication the primacy choice for interaction between the citizens and public services. Identity management makes the foundation for secure and trusted communication, and government frameworks for authentication and identity assurance are therefore developed to support the strategies. This paper examines three existing authentication and identity assurance frameworks, and is a good example to show the importance of specifying assurance frameworks that takes a holistic view of the identity management lifecycle and related threats.

1 Introduction

A (digital) identity is *the information used to represent an entity in an ICT system* [4]. In the context of this paper we think of entity as a human being, meaning that we think of identity as a digital representation of a physical person. A digital identity consist of three key elements [6]: 1) an *identifier* used to identify the owner of the identity 2) *attributes*, which describes different characteristics of, or related to, the identity owner 3) *credentials* which is evidence/data that is used by the identity owner to establish confidence that the person using the identity in the digital world corresponds to the claimed person. There must be processes in place to create, use, update, and revoke digital identities, and policies must exist to govern each of these activities. This is called Identity Management (IdM), and the IdM lifecycle is illustrated in Figure 1. The rigor and quality of all steps of the IdM process can vary substantially between different organizations, and this affects the level of trust that can be associated with a digital identity. Dishonest individuals can exploit weaknesses in any of the identity management lifecycle steps to gain unauthorized access to resources, and as such threaten confidentiality, integrity and availability of assets.

Security requirements can be specified for each phase and each activity in the IdM lifecycle to mitigate threats towards it. The purpose of defining security requirements in relation to identity management is to increase the confidence in the identity establishment phase, and increase the confidence that the individual who uses a digital identity is the individual to whom it was issued [7].

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 343–352, 2013.

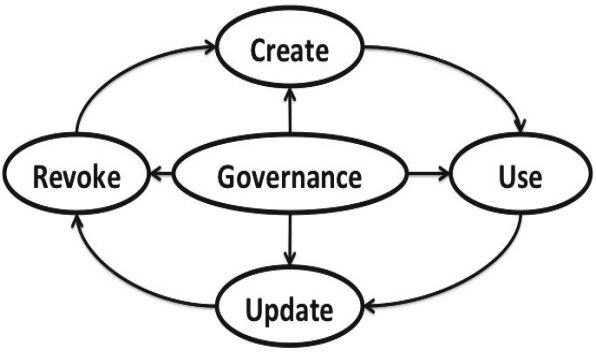


Fig. 1. Identity Management Lifecycle. Adapted from [6]

Requirements for each lifecycle activity can be bundled to form identity assurance levels, where a low assurance level specifies IdM requirements for systems with limited risk levels and high assurance levels define IdM protection strategies in high-risk environments. Examples of assurance levels with associated requirements targeted at the activities in the IdM lifecycle can be found in Identity Assurance Frameworks, such as those defined by the Norwegian government [1], the Australian government [2], and the US government [7].

In this paper we will look at each step of the identity management lifecycle, and identify relevant threats to each lifecycle phase (section 2). The government frameworks mentioned above [1] [2] [7] are examined to determine whether they specify security requirements that can mitigate the identified threats, and they are used in this paper to illustrate the need for holistic identity assurance frameworks that cover all phases of the IdM lifecycle (section 3). Then we provide a discussion of our findings in section 4, and conclude the paper in section 5.

2 Identity Management Lifecycle and Threats towards It

Identity management life cycles have been presented in different shapes, for instance in by International Standards Organization [4], Baldwin et. al [5] and Bertino and Takahashi [6]. Even though the lifecycle presentations vary between these, they treat the same concepts. The following structure, which is illustrated in Figure 1 is inspired by Bertino and Takahashi. More information about threats towards IdM can be found in [7] and [5], while more technical insight to most threats can be found in the CAPEC database[1].

2.1 Creation

The first phase in the IdM lifecycle is identity creation. Identity attributes will be collected and registered, credentials will be defined, and finally issued to the

[1] CAPEC, Common Attack Pattern Enumeration and Classification, `http://capec.mitre.org/`

user during this process. Identity proofing including screening and vetting of users [4] can be part of these activities. The creation process is the foundation for all subsequent use of digital identities, and as such rigor in this phase is of utmost importance for systems that require a high to moderate security level.

Threats to the Creation Process. There are numerous motives for attackers to somehow manipulate the identity creation process, and where one example is to assume the identity of another person during the establishment of a digital identity. This can e.g. be done by presenting forged identity information (e.g. false passport) during the identity proofing process, or exploit the fact that identity proofing is not operationalized in the creation process. University enrollment under a fake alias, establishment of credit cards or establishment of phone subscriptions in another persons name are examples of this threat. The consequence of this is that the attacker obtains full access to resources by means of a valid user identity. Further, invalid attributes can be inserted in the user database, attributes can be modified by unauthorized entities or valid, and false attributes can be registered during the attribute registration if proper countermeasures against these threats are not in place. These threats can have serious consequences knowing that attributes can be used to determine access level e.g. based on group memberships/roles in role based access control (RBAC) schemes or possibly any other attribute in attribute-based access control (ABAC) schemes. Also the credential registration process must be protected so that attackers cannot steal or copy credentials, such as username password pairs. If attackers get access to valid credentials, they can impersonate valid users to obtain protected information. These challenges also exist during delivery of digital identities. Attackers can obtain access to digital identities, which can be used in subsequent malign activities by intercepting the communication channel used to deliver the credentials, such as mail or e-mail.

2.2 Usage

Once a digital identity is created and issued, it is time to start using it in electronic transactions. Digital identities are often being associated with the authentication process. The issued credentials are being used for this purpose. It is also becoming more and more common that electronic services provide personalized content based on identity attributes, and even to base access control decisions on certain identity attributes. The use of digital identities can vary from use on one separate service, to use on multiple services. Single-sign-on (SSO) is a concept where users obtain a security assertion after a successful authentication, and where this assertion is used as authentication evidence towards the subsequent services the user visits. SSO is commonly used in enterprise networks where employees' authentication provides them a security assertion (e.g. Kerberos tickets in Microsoft environments) that can be used to access their e-mail, file shares, intranet and so on. Federated single-sign-on is an extension of the SSO concept, where organizations can cooperate on technology, processes and

policies for identity management. Federated SSO allows security tokens to be used to achieve single-sign-on across organizational borders.

Threats to the Use Phase. There are many threats towards the use of digital identities. Access credentials can be lost, stolen or cracked so that attackers can authenticate, and thereby impersonate, valid users. There are many attack vectors used to obtain valid credentials. Communication lines can be intercepted to copy plaintext data, password files can be stolen and decrypted, social engineering can be used to trick users into giving away their credentials, and so on. The introduction of SSO and federated SSO has added to this complexity in that security assertions are issued based upon a successful authentication. This security assertion is stored by the client and used as proof of identity in subsequent service request. This means that an attacker can copy assertions and add them to malign service requests, or replay previously sent messages. If the receiving service trusts the assertions it will provide information as requested. Since authentication data (assertions) are shared across services in SSO and across services within different trust domains in federated SSO, the attack surface in weakly designed systems is highly increased compared to having separate systems. As already mentioned, RBAC- and ABAC-models allow taking access control decisions based on identity attributes. If attackers can modify attributes during transmission, they can be allowed to elevate their privileges by manipulating attributes. Another scenario is that attackers modify e.g. shipping address so that one user orders and pays the goods, which are then sent to the attacker's destination. The disclosure of identity attributes may also violate users privacy, or reveal company internal information.

2.3 Update

While some identity attributes are static, such as date of birth, eye color and hight, others can change over time. Employees' role in a company can change, people can move and change address, and credit cards, digital certificates and so on can expire. The identity management process must therefore include good procedures to keep identity attributes up to date to ensure their correctness. Identity adjustment, reactivation, maintenance, archive and restore are activities part of the identity update process [4].

Threats to the Update Phase. The threats related to the update phase are similar to those presented in the creation phase. Credentials can be copied or stolen and false attributes can be provided. In operative environments one can experience that the responsibility for identity creation and identity update are placed at different levels in the organization. While the human resource department may be responsible for creation of user identities e.g. in relation with a new employment, the responsibility for updating user profiles may lie at the IT-support department. Consequently, attackers can approach different parts of an organization to achieve the same goals. Attackers can also exploit

weaknesses specific to the update procedures. Delays in the update procedure can allow users to access content based on old but still valid access credentials and attributes, and attacks towards update management interfaces can allow unauthorized reactivation of user accounts.

2.4 Revocation

Identities, including credentials should be revoked if they become obsolete and/or invalid [6]. Revocation can be separated into identity attribute suspension and identity deletion [4]. The former means that some or all identity attributes are made unavailable so that access rights associated with these attributes are made temporarily unavailable to the user. An example of this can be that the association between a user and a certain group membership is removed to reduce a user's access rights. Another is the deactivation of all access rights associated with a user. Identity deletion means the complete removal of registered identity information. Information about revocation should be distributed to all relevant stakeholders to ensure that access is not given based on invalid credentials.

Threats to the Revocation Phase. Suspension and deletion of identity information can primarily be misused to block authorized users from accessing resources. Additionally, insufficient distribution of revocation lists to distributed services can allow attackers to use stolen identities even after the access rights have been formally revoked.

2.5 Governance

There is a need to have policies in place and govern all steps of the identity management lifecycle. Regarding creation of identities, for instance, there should be policies in place that regulate e.g. who can create identities, how they are created, how the quality of attributes can be assured, how credentials are issued and so on. Identity management governance is closely related to identity assurance and identity assurance levels, where requirements for all phases are specified.

Threats to Identity Management Governance. Password policies are among the policies that affect all phases of the identity management lifecycle, so we continue to use this as an example to illustrate the lack of, or weak, policies. Password policies should include requirements for password length, complexity and validity period. Non-existent or weak policies will allow users to associate their digital identities with insecure passwords. Weak passwords are easily being hacked e.g. through brute force attacks or guessing attacks. Insufficient password policies therefore lead to concerns whether an identity can be trusted or not. Non-existent or poor requirements for password change (update) and revocation also affect the trustworthiness of credentials. With infinite password lifetime, attackers can exploit compromised credentials as long as the user account is active. Policy incompliance means that policies exist, but that they are

not being followed to e.g. due to lack of policy enforcement. It does not help to have password length and complexity requirements if the technical platform still allows users to select shorter and weaker passwords. Further, many users will continue to reuse their passwords after expiry, despite a policy stating that passwords are valid for 90 days and that reuse is not allowed. Lack of policies in other IdM areas will similarly lead to weaknesses that can be exploited.

3 Identity Assurance Frameworks

The previous section introduced the steps of the IdM lifecycle and threats that are relevant to each of these. Baldwin et al. [5] state that identity assurance is *concerned with the proper management of risks associated with identity management.* Identity assurance contributes to ensure *confidence in the vetting process used to establish the identity of the individual to whom the credential was issued, and confidence that the individual who uses the credential is the individual to whom the credential was issued* [7]. Identity assurance frameworks consider the threats associated with each IdM lifecycle phase, and specify security requirements to mitigate them.

Many governments around the world, including the Norwegian, the Australian and the US, have developed government strategies to provide online services to their citizens, and to make electronic communication between citizens and the public services a primary choice. There are several legal requirements that regulate such communication, and proper identity management and proper management of identity assurance levels are essential to fulfill them. Consequently, each of these governments have developed identity assurance frameworks: The Norwegian Framework for Authentication and Non-repudiation with and within the Public Sector (FANR) [1], the Australian National e-Authentication Framework (NeAF) [2] and the US National Institute of Standards and Technology (NIST) Electronic Authentication Guideline [7].

Security requirements for each IdM lifecycle phase are bundled to form identity assurance levels; the higher the assurance level, the stricter requirements. Assurance levels can be seen as the levels of trust associated with a credential [9], and information about the assurance level of a digital identity can be used by service providers to determine whether they trust the identity presented to them or not. The US government, for instance, defines four identity assurance levels [3]:

- Level 1: Little or no confidence in the asserted identity's validity
- Level 2: Some confidence in the asserted identity's validity
- Level 3: High confidence in the asserted identity's validity
- Level 4: Very high confidence in the asserted identity's validity

Identities that fulfill requirements at level 1 can be used to access content that has limited concerns regarding confidentiality, integrity and availability, while identities fulfilling level 4 requirements can be used to access assets at the highest classification level. This will balance needs for usability and security.

In Table 1 we provide a summary of the IdM lifecycle phases and activities we presented in section 2, and a third column to illustrate which of the lifecycle phases and activities each assurance framework cover[2]. Our claim is that identity assurance frameworks should cover all phases, and all important activities of the IdM lifecycle to establish trustworthy IdM. Non-existence of requirements may lead to situations where identity risks are not being properly managed.

Table 1. IdM Lifecycle and assurance framework coverage

IdM Lifecycle Phase	**Lifecycle activity**	**Framework coverage**		
		FANR	NeAF	NIST
Create	Credential delivery	x	x	x
	Identity proofing		x	x
	Attribute registration			x
Use	Authentication	x	x	x
	Use of assertions (SSO/federated SSO)			x
	Attribute sharing			
Update	Renew credential		x	x
	Update attributes			
	Reactivate user account		x	x
Revoke	Suspend attributes		x	
	Delete identity		x	x
	Distribute revocation lists		x	x

4 Discussion

As Table 1 illustrates, the most extensive assurance framework of the three we have investigated is the NIST Electronic Authentication Guideline. Both the Australian and the Norwegian frameworks have shortage of requirements for several of the IdM lifecycle activities. Madsen and Itoh [8] state that if there are factors in one lifecycle activity causing low assurance, then this will determine the total assurance level, even if other areas are fully covered at higher assurance levels. In practice this means that even if services offered e.g. by the Norwegian government use authentication mechanisms that satisfy assurance level 4, the full service should be considered to satisfy assurance level 1 at best, since there are no requirements for use of SSO assertions (online services offered by the Norwegian government use federated single-sign-on). We will primarily use the Norwegian framework [1] as example in the following discussion.

The Norwegian framework specifies requirements for the creation phase only targeted at credential delivery. Consequently, threats towards the other activities in the creation phase will not be mitigated unless the identity providers

[2] An x indicates that the framework includes requirements for the given activity, however, the completeness and quality of the requirements are not considered.

implement security controls specified outside the scope of the framework. There are theoretical possibilities that false identity attributes can be registered for a person, and that identities are created for persons with false aliases and so on since there are no common rules for identity proofing and attribute registration. One can also question the quality of created credentials if there are no further specifications regarding credential generation, including key lengths, password strengths and the like.

For the use phase there are requirements targeted at authentication activity. In isolation, the authentication requirements in the Norwegian framework seems to be sufficient in that the quality of the authentication mechanisms shall improve with increasing assurance levels. However, since the identity proofing and other credential quality requirements during the creation phase are not in place there is still a risk that credentials in use are of low quality, and therefore exposed to guessing attacks or brute force attacks. Further, the framework does not specify any protection requirements for use of assertions. If the assertions in SSO and federated SSO environments are not properly protected, an attacker can intercept the communication between a user and a public service, copy the assertion, and craft his own service requests with valid assertions included. In this way an attacker can impersonate a user without a need to know access credentials. None of the three investigated identity assurance frameworks specify requirements for the attribute sharing activity. Thomas and Meinel [10] claim that *a verification of an attribute might not be desired as long as a user is not involved in transactions that require it.* As such, the lack of attribute sharing requirements may indicate that there is only a very limited set of attributes being shared in the government systems and that attributes are not being used as source for authorization decisions. If this is not true, however, Thomas and Meinel's advice to implement mechanisms to verify the quality and integrity of shared identity attributes should be followed [10].

Both the Australian (NeAF) and US (NIST) frameworks cover important aspects of the identity update and revocation phases, except that they do not specify requirements on updating and suspending attributes. The reason for omitting such requirements may be similar to what we identified for attribute sharing in the use phase. The Norwegian framework, on the other hand, fails to target the update and revocation phases at large. Users of the Norwegian framework must therefore on an individual basis define security controls to mitigate the threats against the update and revocation phases.

All the government frameworks are developed to facilitate common identity management practices throughout government agencies, and reuse of authentication services or access credentials across online services offered by the governments. Based on the discussion above one can argue that this goal can be fulfilled by following NeAF and NIST guidelines. The Norwegian identity assurance framework [1], on the other hand, has considerable limitations. The Norwegian framework states that *"the factors used to separate between security levels* [read: assurance levels] *are not exhaustive."*. This understatement is consistent with our analysis that shows there are many factors that are not considered at

all. The consequence is that service providers independently need to fill in the gaps where the framework is incomplete. The probability that two independent organizations solves this task completely different is high. There are at least two challenges related to this:

- Specifications, policies and technical solutions will likely be inconsistent. This will result in lack of interoperability between systems, and thus prevent reuse of solutions.
- Each organization will specify different requirements and policies for each assurance level. It will be difficult to assess the assurance level against trustworthiness of the digital identities if there are no common definitions of what each assurance level include.

Madsen and Itoh [8] took at technical view to explain challenges related to identity assurance, and related technical interoperability issues. Our results show that challenges with identity assurance can be elevated to a higher level if identity assurance frameworks are not developed with an holistic view on the identity management lifecycle, i.e. it must be developed to include security requirements that mitigate current threats towards each lifecycle phase. The trust an entity will associate with a digital identity will depend on *all the processes, technologies, and protections followed by the identity provider and on which the digital identity were based* [8]. That being said, the Norwegian Government and public administrations have had success with implementation of a common authentication service for the public sector. The main reason for this is that one common entity, the Agency for Public Management and eGovernment (Difi)[3], has been responsible for realization of a public authentication service (MinID/ID-porten). Norwegian public administrations can integrate their online services with this common authentication portal. The chance of having interoperable, federated SSO enabled, authentication services without this model would have been low without considerable efforts to improve the common Norwegian identity assurance framework, or without substantial coordination activities between the public services.

5 Conclusion

The essence of information security is to protect confidentiality, integrity and availability of assets. To achieve this we need to know whether the entity requesting an asset is authorized or not, and consequently we need to determine the identity of the requestor. Identity management defines the processes and policies to create, use, update and revoke digital identities. IdM is as such essential to ensure information security. Identity assurance frameworks specify requirements targeting the different phases of the identity management lifecycle, and are intended to specify and determine the trustworthiness of digital identities.

[3] `www.difi.no`

In this paper we studied the Norwegian Framework for Authentication and Non-repudiation in Electronic Communication with and within the Public sector, the Australian National e-Authentication framework, and the US Electronic Authentication Guideline as examples of existing identity assurance frameworks. We saw that these frameworks have considerable deviations in coverage when it comes to targeting security requirements towards the identity management lifecycle phases and activities. The paper illustrates the importance of specifying assurance frameworks that takes a holistic view of the identity management lifecycle and related threats.

References

1. Framework for autehntication and non-repudiation in electronic communication with and within the public sector (norwegian title: Rammeverk for autentisering og uavviselighet i elektronisk kommunikasjon med og i offentlig sektor. Tech. rep. Det kongelige fornyings og administrasjonsdepartementet, Norwegian Government (2008)
2. National e-authentication framework. Tech. rep. Australian Government, Department of Finance and Deregulation (2009)
3. E-authentication guidance for federal agencies. Tech. Rep. OMB Memorandum M-04-04 (2011)
4. Information technology - security techniques - a framework for identity management - part 1: Terminology and concepts. Tech. Rep. ISO/IEC 24760-1, ISO/IEC (2011)
5. Baldwin, A., Mont, M.C., Shiu, S.: On identity assurance in the presence of federated identity management systems. In: Proceedings of the 2007 ACM Workshop on Digital Identity Management, DIM 2007 (2007)
6. Bertino, E., Takahashi, K.: Identity Management - Concepts, Technologies and Systems. Artech House (2011)
7. Burr, W.E., Dodson, D.F., Newton, E.M., Perlner, R.A., Polk, W.T., Gupta, S., Nabbus, E.A.: Electronic authentication guideline. Tech. Rep. Special Publication 800-63-1, National Institute of Standards and Technology (2011)
8. Madsen, P., Itoh, H.: Challenges to supporting federated assurance. Computer 42(5), 42–49 (2009)
9. Soutar, C., Brenan, J.: Identity assurance framework: Overview. Tech. rep. Kantara initiative (2010)
10. Thomas, I., Meinel, C.: An attribute assurance framework to define and match trust in identity attributes. In: 2011 IEEE International Conference on Web Services, ICWS, pp. 580–587 (July 2011)

Anonymous Lattice-Based Broadcast Encryption*

Adela Georgescu

Faculty of Mathematics and Computer Science, University of Bucharest,
Academiei Street 14, Bucharest 010014, Romania
adela@fmi.unibuc.ro

Abstract. In this paper we propose a lattice-based anonymous broadcast encryption scheme obtained by translating the broadcast encryption scheme of Paterson et al. [7] into the lattices environment. We use two essential cryptographic primitives for our construction: tag-based hint systems secure under Ring-LWE hardness and IND-CCA secure cryptosystem under LWE-hardness. We show that it is feasible to construct anonymous tag-based hint systems from Ring-LWE problem for which we use a variant with "small" secrets known to be as hard as regular Ring-LWE. We employ an IND-CCA-secure public key encryption scheme from LWE [12] for the PKE component of the anonymous broadcast encryption scheme.

Keywords: broadcast encryption, anonymity, Learning With Errors, Lattices.

1 Introduction

In this paper, we translate the anonymous broadcast encryption scheme from [7] into the lattices environment. Lattices are more and more studied recently and lattices environment is becoming wider and more populated with different cryptographic primitives. They offer certain undeniable advantages over traditional cryptography based on number theory: hard problems which form the security basis of many cryptographic primitives, great simplicity involving linear operations on small numbers and increasingly efficient implementations. A very important issue is that they are believed to be secure against quantum attacks in an era where quantum computers are a great promise for the near future. It is not surprising that lately we are witnessing a great development of cryptographic constructions secure under lattice-based assumptions. This is the main motivation for our current work: we want to propose a lattice-based variant of this cryptographic primitive (i.e. anonymous broadcast encryption) existent in classical cryptography.

Authors from [7] use two cryptographic primitives in order to achieve anonymous broadcast encryption: IND-CCA public key encryption scheme and

* This work was sponsored by the European Social Fund, under doctoral and post-doctoral grant POSDRU/88/1.5/S/56668.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 353–362, 2013.

anonymous tag-based hint system. We employ variants of both these primitives derived from the Ring-Learning With Errors problem (RLWE) introduced recently in [10]. This problem is the ring-based variant of Regev's Learning With Errors problem [13]. Lyubashevsky et al. [10] show that their problem can be reduced to the worst-case hardness of short-vector problems in ideal lattices. The advantage of RLWE based cryptographic primitives over LWE-based cryptographic primitives is that they achieve more compact ciphertext and smaller key sizes by a factor of n, thus adding more efficiency.

The RLWE problem has already been used as underlying hardness assumption for many cryptographic constructions, starting with the original cryptosystem from [10] and continuing with efficient signature schemes [9], [12], pseudo-random functions [2], fully homomorphic encryption [3] and also NTRU cryptosystem [14]. So it is a natural question to ask if we can achieve anonymous broadcast encryption from lattices. As one can see in the rest of the paper, we found it is not hard to construct this kind of primitive. IND-CCA cryptosystem based on LWE problem (and also RLWE) were already introduced in the literature (see section 6.3 [12] for LWE-based IND-CCA cryptosystem). We also prove that it is feasible to construct tag-based hint anonymous systems from RLWE following the model of DDH hint system from [7]. For this specific task, we deal with the Hermite Normal Form variant of RLWE and with an equivalent version of DDH problem based on RLWE introduced in [5].

1.1 Related Work

There is another candidate in the literature for lattice-based broadcast encryption scheme introduced in [15]. Anyway, there are some important differences between our scheme and this one: the latter does not offer anonymity but it is an identity-based scheme. Our scheme can also be transformed into identity-based broadcast encryption by replacing the LWE-based IND-CCA secure PKE with identity-based encryption (IBE) from LWE as the one from [4]. On the other hand, the CCA-secure PKE scheme from [12] we employ in our construction has better efficiency and simplicity due to the simple structure of the new trapdoor they introduce, thus also making our construction more efficient.

2 Preliminaries

2.1 Lattices

Let $\mathbf{B} = \{\mathbf{b_1}, ...\mathbf{b_n}\} \in \mathbb{R}^{n \times k}$ be linearly independent vectors in $\mathbb{R}^n$. The lattice generated by $\mathbf{B}$ is the set of all *integer* linear combinations of vectors from $\mathbf{B}$

$$\mathcal{L}(\mathbf{B}) = \{\sum_{i=1}^{n} x_i \cdot \mathbf{b_i} \ : \ x_i \in \mathbb{Z}\}.$$

Matrix $\mathbf{B}$ constitutes a basis of the lattice. Any lattice admits multiple bases, some bases are better than others.

We introduce here a function that we'll apply in section 3.1, the $round(\cdot)$ function. This function was first used with its basic variant in [13] for decryption, and later on to almost all the lattice-based cryptosystems :

$$round(x) = \begin{cases} 1, x \in [0, \lfloor q/2 \rfloor] \\ 0, \text{ otherwise} \end{cases}$$

In our construction, we use the extended variant of the function which rounds to smaller intervals, namely $round(x) = a$ if $x \in [a \cdot q/A, (a+1) \cdot q/A]$ where A is the total number of intervals. We suggest setting $A = 4$.

We employ this function in order to derive the same value from numbers that are separated only by a small difference (Gaussian noise).

2.2 The Learning with Errors Problem

The learning with errors problem (LWE) is a recently introduced (2005, [13]) but very famous problem in the field of lattice-based cryptography. Even if it is not related directly to lattices, the security of many cryptographic primitives in this field rely on its hardness believed to be the same as worst-case lattice problems.

Informally, the problem can be described very easily: given n linear equations on $s \in \mathbb{Z}_q^n$ which have been perturbed by a small amount of noise, recover the secret s.

We present here the original definition from [13].

Definition 1. *(The Learning With Errors Problem [13])*
Fix the parameters of the problem: $n \geq 1$, modulus $q \geq 2$ and Gaussian error probability distribution χ on $\mathbb{Z}_q$ (more precisely, it is chosen to be the normal distribution rounded to the nearest integer, modulo q with standard deviation αq where $\alpha > 0$ is taken to be $1/(poly(n))$). Given an arbitrary number of pairs $(\mathbf{a}, \mathbf{a^T s} + e)$ where s is a secret vector from $\mathbb{Z}_q^n$, vector $\mathbf{a}$ is chosen uniformly at random from $\mathbb{Z}_q^n$ and e is chosen according to χ, output $\mathbf{s}$ with high probability.

Proposition 1. *[13] Let $\alpha = \alpha(n) \in (0,1)$ and let $q = q(n)$ be a prime such that $\alpha q > 2\sqrt{n}$. If there exists an efficient (possibly quantum) algorithm that solves $LWE_{q,\chi}$, then there exists an efficient quantum algorithm for approximating SIVP in the worst-case to within $O(n/\alpha)$ factors.*

2.3 The Ring-Learning with Errors Problem

The *ring learning with errors* assumption introduced by Lyubashevsky et al. [10] is the translation of the LWE into the ring setting. More precisely, the group $\mathbb{Z}_q^n$ from the LWE samples is replaced with the ring $R_q = \mathbb{Z}_q[x]/\langle x^n + 1 \rangle$, where n is a power of 2 and q is a prime modulus satisfying $q = 1 \bmod 2n$. This is in fact a particularization of the ring-LWE problem introduced in the original paper, but for our construction, as for many others, it is enough. The ring $\mathbb{Z}_q[x]/\langle x^n + 1 \rangle$

contains all integer polynomials of degree $n-1$ and coefficients in $\mathbb{Z}_q$. Addition and multiplication in this ring are defined modulo x^n+1 and q.

In ring-LWE [10], the parameter setting is as follows: $s \in R_q$ is a fixed secret, a is chosen uniformly from R_q and e is an error term chosen independently from some error distribution χ concentrated on "small" elements from R_q. The ring-LWE (RLWE) assumption is that it is hard to distinguish samples of the form $(a, b = a \cdot s + e) \in R_q \times R_q$ from samples (a, b) where a, b are chosen uniformly in R_q. A hardness result based on the worst-case hardness of short-vector problems on ideal lattices is given in [10]. An important remark is that the assumption still holds if the secret s is sampled from the noise distribution χ rather than the uniform distribution; this is the "Hermite Normal Form (HNF)" of the assumption (HNF-ring-LWE). The advantage of the RLWE problem is that it represents a step forward in making the lattice-based cryptography practical. In most applications, a sample $(a, b) \in R_q \times R_q$ from RLWE distribution can replace n samples $(\mathbf{a}, b) \in \mathbb{Z}_q^n \times \mathbb{Z}_q$ from the standard LWE distribution, thus reducing the key size by a factor of n.

We note that in our construction of the broadcast encryption scheme, we will make use of the HNF form of the RLWE problem.

We present in the following the correspondent of the Decisional Diffie-Hellman based on the Ring-LWE problem, which was first introduced in [5] and which is derived from the ring-LWE cryptosystem from [8], section 3.1. The security of this cryptosystem is proven conditioned by the fact that an adversary cannot solve the below problem, which is essentially its view from the cryptosystem.

DDH-RLWE Problem. [5] Given a tuple $(s, y_1 = s \cdot x + e_x, y_2 = s \cdot y + e_y, z)$ where s is chosen uniformly at random from R_q, x, y, e_x, e_y are sampled from χ distribution, one has to distinguish between the tuple where $z = y_1 \cdot y + e_3$, with e_3 sampled independently from χ and the same tuple where z is chosen uniformly and independently from anything else in R_q.

We present a hardness result for the above problem but, due to lack of space, we defer a complete proof to [5].

Proposition 2. *[5]*
The DDH-RLWE problem is hard if the RLWE problem in its "Hermite normal form" (HNF) is hard.

3 Anonymous Broadcast Encryption

In this section we recall a general Broadcast Encryption model from [7] which allows anonymity.

Definition 2. *A broadcast encryption scheme with security parameter λ and $U = \{1, ..., n\}$ the universe of users consists of the following algorithms.*

`Setup`(λ, n) *takes as input security parameter λ and the number of users and outputs a master public key MPK and master secret key MSK.*

`KeyGen`(MPK, MSK, i) *takes as input MPK, MSK and* $i \in U$ *and outputs the private key* sk_i *corresponding to user* i.

`Enc`(MPK, m, S) *takes as input MPK and a message* m *to be broadcasted to a set of users* $S \subseteq U$ *and it outputs a cipheretxt* c.

`Dec`(MPK, sk_i, c) *takes as input MPK, a private key* sk_i *and a ciphertext* c *and outputs either the message* m *or a failure symbol.*

We provide the same security model as in [7] for the anonymous broadcast encryption scheme we'll describe later.

Definition 3. *We define the ANO-IND-CCA security game (against adaptive adversaries) for broadcast encryption scheme as follows.*

Setup. *The challenger runs the Setup to generate the public key MPK and the corresponding private key MSK and gives MPK to the adversary* $\mathcal{A}$.

Phase 1. $\mathcal{A}$ *can issues two types of queries:*

- *private key extraction queries to an oracle for any index* $i \in U$*; the oracle will respond by returning the private key* $sk_i = KeyGen(MPK, MSK, i)$ *corresponding to* i*;*
- *decryption queries* (c, i) *to an oracle for any index* $i \in U$*; the oracle will respond by returning the* $Dec(MPK, sk_i, c)$.

Challenge. *The adversary selects two equal length messages* m_0 *and* m_1 *and two distinct sets* S_0 *and* $S_1 \subseteq U$ *of users. We impose the same requirements as in [7]: sets* S_0 *and* S_1 *should be of equal size and A has not issued any query to any* $i \in (S_0 \setminus S_1) \cup (S_1 \setminus S_0)$*. Further, if there exists an* $i \in S_0 \cap S_1$ *for which A has issued a query, then we require that* $m_0 = m_1$*. The adversary gives* m_0*,* m_1 *and* S_0*,* S_1 *to the challenger. The latter picks a random bit* $b \in \{0, 1\}$*, computes* $c^* = Enc(MPK, m_b, S_b)$ *and returns it to A.*

Phase 2. $\mathcal{A}$ *continues to issue private key extraction queries with the restriction that* $i \notin (S_0 \setminus S_1) \cup (S_1 \setminus S_0)$*; otherwise it is necessary that* $m_0 = m_1$*.* $\mathcal{A}$ *continues to issue decryption queries* (c, i) *with the restriction that if* $c = c^*$ *then either* $i \notin (S_0 \setminus S_1) \cup (S_1 \setminus S_0)$ *or* $i \in S_0 \cap S_1$ *and* $m_0 = m_1$.

Guess. *The adversary A outputs a guess* $b' \in \{0, 1\}$ *and wins the game if* $b = b'$.

We denote $\mathcal{A}'s$ advantage by $Adv_{A,KT}^{ANO-IND-CPA}(\lambda) = |Pr[b' = b] - \frac{1}{2}|$ where λ is the security parameter of the scheme.

Generic constructions for anonymous broadcast encryption can be obtained exactly as in Section 3 and 4 from [7], but they require linear time decryption. Thus, we follow the idea of introducing tag-based anonymous hint system as in [7], but we construct it from the ring-LWE problem. The construction has the advantage of achieving constant time decryption.

3.1 Tag-Based Anonymous Hint Systems

A tag-based anonymous hint system (TAHS) [7] is a sort of encryption under a tag t and a public key pk. The output is a pair (U, H) where H is a hint. This pair should be hard to distinguish when using two different public keys. Such a system consists of the following algorithms:

$\texttt{KeyGen}(\lambda)$ on input security parameter λ, outputs a key pair (sk, pk).

$\texttt{Hint}(t, pk, r)$ takes as input a public key pk and a tag t; outputs a pair (U, H) consisting of a value U and a hint H. It is required that U depends only on random r and not on pk.

$\texttt{Invert}(sk, t, U)$ takes as input a value U, a tag t and a private key sk. It outputs either a hint H or $\perp$ if U is not in the appropriate domain.

Correctness implies that for any pair $(sk, pk) \leftarrow \texttt{KeyGen}(\lambda)$ and any random r, if $(U, H) \leftarrow \texttt{Hint}(t, pk, r)$, then $\texttt{Invert}(sk, t, U) = H$.

Definition 4. *[7]*
A tag-based hint system as defined above is anonymous if there is no polynomial time adversary which has non-negligible advantage in the following game:

1. *Adversary $\mathcal{A}$ chooses a tag t' and sends it to the challenger.*
2. *The challenger generates two pairs $(sk_0, pk_0), (sk_1, pk_1) \leftarrow \texttt{KeyGen}(\lambda)$ and gives pk_0, pk_1 to the adversary.*
3. *The following phase is repeated polynomially many times: $\mathcal{A}$ invokes a verification oracle on a value-hint-tag triple (U, H, t) such that $t \neq t'$. In reply, the challenger returns bits $d_0, d_1 \in \{0, 1\}$ where $d_0 = 1$ if and only if $H = \texttt{Invert}(sk_0, t, U)$ and $d_1 = 1$ if and only if $H = \texttt{Invert}(sk_1, t, U)$.*
4. *In the challenge phase, the challenger chooses random bit $b \leftarrow \{0, 1\}$ and random $r' \leftarrow R_q$ and outputs $(U', H') = \texttt{Hint}(t', pk_b, r')$.*
5. *$\mathcal{A}$ is allowed to make any further query but not involving target t'.*
6. *$\mathcal{A}$ outputs a bit $b' \in \{0, 1\}$ and wins the game if $b' = b$.*

To show that this primitive can be constructed in the lattice-based environment, we give an example of an anonymous hint system based on the DDH-RLWE assumption. This is the equivalent of the hint system based on the classical DDH assumption from [7].

Let R_q be the ring of polynomial integers as described in section 2.3 i.e. $R_q = \mathbb{Z}_q^n / \langle x^n + 1 \rangle$ where n is a power of 2 and q is a prime modulus such that $q = 1 \bmod 2n$. Remember that χ is the noise distribution concentrated on "small" elements from R_q; s is a fixed element from R_q.

We draw attention to the fact that, unlike in the tag-based hint system from [7], the `Hint` algorithm outputs a value H_1 which is slightly different from the value H_2 recovered by `Invert` algorithm (by a small quantity from χ as shown below) and only the holder of the secret key sk can derive a value H from both H_1 and H_2. We stress that the final value H is the same for every use of the tag-based hint scheme, just that is somehow hidden by the output of `Hint` algorithm.

$\texttt{KeyGen}(\lambda)$ take random $x_1, x_2, y_1, y_2, e_1, e_2, e_1', e_2' \leftarrow \chi$ and compute $X_i = s \cdot x_i + e_i$ and $Y_i = s \cdot y_i + e_i'$. The public key is $pk = (X_1, X_2, Y_1, Y_2)$ and the private key is $sk = (x_1, x_2, y_1, y_2)$.

$\texttt{Hint}(t, pk, r)$ choose e, e_x, e_y from χ distribution and compute (U, H_1) as

$$U = s \cdot r + e; \quad H_1 = (V, W) = ((X_1 \cdot t + e_x + X_2)r, (Y_1 \cdot t + e_y + Y_2)r)$$

`Invert`(sk, t, U) parse sk as (x_1, x_2, y_1, y_2), compute

$$H_2 = (V, W) = (U(t \cdot x_1 + x_2), U(t \cdot y_1 + y_2))$$

and then check if the difference $H_2 - H_1$ is small (i.e. from χ distribution). If this is true, then output
$round(H_2) = (round(U(t \cdot x_1 + x_2)), round(U(t \cdot y_1 + y_2))) = round(H_1) = H$

Let us now check the correctness of the scheme. We note that the output of `Hint` algorithm is the pair (U, H_1) where $U = s \cdot r + e$. After some simplifications, we obtain

$H_1 = (s \cdot r \cdot (t \cdot x_1 + x_2) + (e_1 \cdot t + e_x + e_2) \cdot r, s \cdot r \cdot (t \cdot y_1 + y_2) + (e'_1 \cdot t + e_y + e'_2) \cdot r)$
where $(e_1 \cdot t + e_x + e_2) \cdot r$ and $(e'_1 \cdot t + e_y + e'_2) \cdot r$ are "small" since they both belong to the χ distribution.

On the other hand, H_2 will be computed as
$H_2 = (s \cdot r \cdot (t \cdot x_1 + x_2) + (t \cdot x_1 + x_2) \cdot e, s \cdot r \cdot (t \cdot y_1 + y_2) + (t \cdot y_1 + y_2) \cdot e)$
again with $(t \cdot x_1 + x_2) \cdot e$ and $(t \cdot y_1 + y_2) \cdot e$ both small from χ.

Therefore, the difference $H_2 - H_1$ is small and belongs to χ. Thus, by computing both $round(H_1)$ and $round(H_2)$, one gets exactly the same value, which is in fact hidden in the output of `Hint` algorithm.

Lemma 1. *The above tag-based hint system is anonymous if the DDH-RLWE assumption holds in the ring R_q.*

Proof. The proof of this lemma follows closely that of Lemma 1 from [7] adapted to the LWE environment. We will give a sketch of it in the following.

The proof is modeled by a sequence of games, starting with the first game which is the real game.

Game 0 is the real attack game.

Game 1 differs from Game 0 in the following two issues: the challenger's bit b is chosen at the beginning of the game and in the adversary's challenge $(U^*, (V^*, W^*))$, W^* is replaced by a random element of R_q.
We show that a computationally bounded adversary cannot distinguish the adversary's challenge $(U^*, (V^*, W^*))$ from the one where W^* is replaced by a random element from R_q, under the DDH-RLWE assumption.
We construct a DDH-RLWE distinguisher B for Game 0 and Game 1 which takes as input $(s, X = s \cdot x + e_x, Y = s \cdot y + e_y, Z)$ where x, y, e_x, e_y are from χ and aims at distinguishing whether $Z = X \cdot y + e_z$ or Z is random in R_q. At the beginning of the game, B chooses θ_1 and θ_2 from χ and defines $X' = s \cdot \theta_1 + X \cdot \theta_2$. When the challenge bit b is chosen, B generates pk_{1-b} by choosing $x_{1-b,1}, x_{1-b,2}, y_{1-b,1}, y_{1-b,2}, e_{1-b,1}, e_{1-b,2}, e_{1-b,1}, e_{1-b,2} \leftarrow \chi$ and setting $X_{1-b,i} = s \cdot x_{1-b,i} + e_{1-b,i}$, for $i \in \{1, 2\}$. For pk_b, B chooses $\alpha, \beta_1, \beta_2 \leftarrow \chi$ and computes $X_{b,1} = X', X_{b,2} = X' \cdot (-t^*) + s \cdot \beta_1, Y_{b,1} = s \cdot \beta_2 + X \cdot \alpha$ and $Y_{b,2} = s \cdot (-\beta_2) \cdot t^*$. The adversary is given the public keys $(X_{0,1}, X_{0,2}, Y_{0,1}, Y_{0,1})$ and $(X_{1,1}, X_{1,2}, Y_{1,1}, Y_{1,1})$.

To answer a verification query $(U,(V,W),t)$ with $t \neq t^*$ coming from adversary A, B can run algorithm $\texttt{Invert}(sk_{1-b},t,U)$ since he knows sk_{1-b}. As for $\texttt{Invert}(sk_b,t,U)$, he computes
$Z_1 = (V - U \cdot \beta_1) \cdot 1/(t-t^*)$ $Z_2 = (w - U \cdot \beta_2(t-t^*)) \cdot 1/\alpha t$
and answers that $d_b = 1$ if and only if $round(Z_1) = round(U \cdot \theta_1 \cdot Z_2 \cdot \theta_2)$.
First of all, we note that we are working in the ring $\mathbb{Z}_q^n / \langle x^n + 1 \rangle$ which is a field , since q is prime and $x^n + 1$ is irreducible. Therefore, the multiplicative inverse is defined and we can compute $(1/(t-t^*))$ for example.
Finally, in the challenge phase, B constructs the challenge pair $(U^*,(V^*,W^*))$ as $U^* = Y,\ V^* = Y \cdot \beta_1,\ W^* = T \cdot \alpha t^*$. If $T = X \cdot y + e_{xy}$ with $e_{xy} \leftarrow R_q$, then A's view is the same as in Game 0 (except with small probability) while if T is random in R_q A's view is the same as in Game 1. Therefore, we have $|Pr[S_1] - Pr[S_0]| \leq Adv^{DDH}(B) + "small"$.

Game 2 is identical to Game 1 but in the challenge phase both V^* and W^* are chosen uniformly in R_q and independent of U^*. We argue that adversary A cannot see the difference as long as the DDH-RLWE assumption holds. In this game, the challenge is just a sequence of random ring elements and we have $Pr[S_2] = 1/2$.
By combing the above informations, we obtain

$$Adv^{anon-hint}(A) \leq 2Adv^{DDH}(B) + 2q/p.$$

3.2 Anonymous Broadcast Encryption

In this subsection we construct the anonymous broadcast encryption scheme from anonymous hint system $S_{hint} = (\texttt{KeyGen}, \texttt{Hint}, \texttt{Invert})$ based on LWE and LWE-based public key encryption scheme $S_{pke} = (\texttt{Gen}, \texttt{KeyGen}, \texttt{Encrypt}, \texttt{Decrypt})$. We also need a LWE-based signature scheme $\Sigma = (\mathcal{G}, \mathcal{S}, \mathcal{V})$. We remark that this is precisely the construction from [7], since in this stage of description, we don't have any contribution to it. Our contribution was mainly to translate the TAHS scheme in the lattice-based environment.

$\texttt{Setup}(\lambda, n)$: Obtain $par \leftarrow \texttt{Gen}(\lambda)$ and, for $i = 1$ to n generate encryption key pairs $(sk_i^e, pk_i^e) \leftarrow S^{pke}.\texttt{KeyGen}(par)$ and hint key pairs $(sk_i^h, pk_i^h) \leftarrow S^{hint}.\texttt{KeyGen}(\lambda)$; the master public key consists of

$$MPK = (par, \{pk_i^e, pk_i^h\}_{i=1}^n, \Sigma)$$

and the master secret key is $MSK = \{sk_i^e, sk_i^h\}_{i=1}^n$

$\texttt{KeyGen}(MPK, MSK, i)$: parse $MSK = \{sk_i^e, sk_i^h\}_{i=1}^n$ and output $sk_i = (sk_i^e, sk_i^h)$.

$\texttt{Enc}(MPK, M, S)$: to encrypt a message M for a set of users $S = \{i_1, ..., i_l\} \subseteq \{1, ..., n\}$, generate a signature key pair $(SK, VK) = \mathcal{G}(\lambda)$. Then choose random $r, e \leftarrow \chi$ and compute $(U, H_j) = S^{hint}.\texttt{Hint}(VK, pk_{i_j}^h, r)$ for $j = 1$ to l. Then, for each user index $j \in \{1, ..., l\}$ compute a ciphertext

$C_j = S^{pke}.\texttt{Encrypt}(pk^e_{i_j}, M||VK)$. Choose a random permutation $\pi : \{1, ..., l\} \to \{1, ..., l\}$ and output the final ciphertext as

$$C = (VK, U, (H_{\pi(1)}, C_{\pi(1)}), ..., (H_{\pi(l)}, C_{\pi(l)}), \sigma)$$

where $\sigma = \mathcal{S}(SK, U, (H_{\pi(1)}, C_{\pi(1)}), ..., (H_{\pi(l)}, C_{\pi(l)}))$

$\texttt{Dec}(MPK, sk_i, C)$: for $sk_i = (sk^e_i, sk^h_i)$ and $C = (VK, U, (H_{\pi(1)}, C_{\pi(1)}), ..., (H_{\pi(l)}, C_{\pi(l)}), \sigma)$, return $\perp$ if $\mathcal{V}(VK, U, (H_{\pi(1)}, C_{\pi(1)}), ..., (H_{\pi(l)}, C_{\pi(l)}), \sigma) = 0$ or if U is not in the appropriate space. Otherwise, compute $H = S^{hint}.\texttt{Invert}(sk^h_i, VK, U)$. If $H \neq H_j$ for all $j \in \{1, ..., l\}$, return $\perp$. Otherwise, let j be the smallest index such that $H = H_j$ and compute $M' = S^{pke}.\texttt{Decrypt}(sk^e_i, C_j)$. If M' can be parsed as $M' = M||VK$, return M. Otherwise, return $\perp$.

We already presented an anonymous tag-based hint system secure under Ring-LWE problem. As for the PKE component of the above scheme, we suggest using the IND-CCA secure scheme described in [12]. As the authors claim, it is more efficient and compact than previous lattice-based cryptosystems since it uses a new trapdoor which is simpler, efficient and easy to implement. Under the same reasons, we suggest also employing a lattice-based signature scheme from [12], section 6.2.

Due to lack of space, we can not present any of these two suggested cryptographic primitives here but we refer the reader to [12] for more details. We just mention that they were proven to be secure under LWE-assumption. We note that the tag-based hint system and the PKE cryptosystem employed are independent in our lattice broadcast encryption scheme. Therefore, the fact that the components of the ciphertext are elements from different algebraic structures is not prohibitive. In order to apply the signature scheme, one needs to first apply a hash function on the input with the aim of "smoothing" it.

Theorem 1. *The above broadcast encryption scheme is ANO-IND-CCA secure assuming that S^{hint} scheme is anonymous, the S^{pke} scheme is IND-CCA secure and the signature scheme Σ is strongly unforgeable.*

We remark that the proof of Theorem 4 from [7] is also valid for our theorem since it deals with general IND-CCA encryption scheme and tag-based hint systems, and not with some specific constructions in a certain environment (like traditional cryptography or lattice-based cryptography).

4 Conclusions

We introduced a lattice-based variant of the anonymous broadcast encryption scheme from [7]. We showed that it is feasible to construct anonymous tag-based hint scheme from the RLWE assumption in order to achieve anonymity of the scheme. We used a variant of RLWE assumption with "small" secrets and proved that the hint scheme is anonymous based on a RLWE-based DDH assumption. For public key encryption, we suggested the use of the IND-CCA secure LWE-based encryption scheme and digital signature scheme from [12] as they gain in efficiency and simplicity over the previous similar constructions from lattices.

References

1. Ajtai, M.: Generating Hard Instances of the Short Basis Problem. In: Wiedermann, J., Van Emde Boas, P., Nielsen, M. (eds.) ICALP 1999. LNCS, vol. 1644, pp. 1–9. Springer, Heidelberg (1999)
2. Banerjee, A., Peikert, C., Rosen, A.: Pseudorandom Functions and Lattices. In: Pointcheval, D., Johansson, T. (eds.) EUROCRYPT 2012. LNCS, vol. 7237, pp. 719–737. Springer, Heidelberg (2012)
3. Brakerski, Z., Vaikuntanathan, V.: Fully Homomorphic Encryption from Ring-LWE and Security for Key Dependent Messages. In: Rogaway, P. (ed.) CRYPTO 2011. LNCS, vol. 6841, pp. 505–524. Springer, Heidelberg (2011)
4. Gentry, C., Peikert, C., Vaikuntanathan, V.: Trapdoors for hard lattices and new cryptographic constructions. In: 40th Annual ACM Symposium on Theory of Computing, pp. 197–206. ACM, New York (2008)
5. Georgescu, A., Steinfeld, R.: Lattice-based key agreement protocols. In: Preparation (2013)
6. Gentry, C., Waters, B.: Adaptive Security in Broadcast Encryption Systems (with Short Ciphertexts). In: Joux, A. (ed.) EUROCRYPT 2009. LNCS, vol. 5479, pp. 171–188. Springer, Heidelberg (2009)
7. Libert, B., Paterson, K., Quaglia, E.: Anonymous Broadcast Encryption: Adaptive Security and Efficient Constructions in the Standard Model. In: Fischlin, M., Buchmann, J., Manulis, M. (eds.) PKC 2012. LNCS, vol. 7293, pp. 206–224. Springer, Heidelberg (2012)
8. Lindner, R., Peikert, C.: Better Key Sizes (and Attacks) for LWE-Based Encryption. In: Kiayias, A. (ed.) CT-RSA 2011. LNCS, vol. 6558, pp. 319–339. Springer, Heidelberg (2011)
9. Lyubashevsky, V.: Lattice Signatures without Trapdoors. In: Pointcheval, D., Johansson, T. (eds.) EUROCRYPT 2012. LNCS, vol. 7237, pp. 738–755. Springer, Heidelberg (2012)
10. Lyubashevsky, V., Peikert, C., Regev, O.: On Ideal Lattices and Learning with Errors over Rings. In: Gilbert, H. (ed.) EUROCRYPT 2010. LNCS, vol. 6110, pp. 1–23. Springer, Heidelberg (2010)
11. Micciancio, D., Goldwasser, S.: Complexity of Lattice Problems: a cryptographic perspective. Kluwer Academic Publishers, Boston (2002)
12. Micciancio, D., Peikert, C.: Trapdoors for Lattices: Simpler, Tighter, Faster, Smaller. In: Pointcheval, D., Johansson, T. (eds.) EUROCRYPT 2012. LNCS, vol. 7237, pp. 700–718. Springer, Heidelberg (2012)
13. Regev, O.: On lattices, learning with errors, random linear codes, and cryptography. In: 37th Annual ACM Symposium on Theory of Computing, pp. 84–93. ACM, New York (2005)
14. Stehlé, D., Steinfeld, R.: Making NTRU as secure as worst-case problems over ideal lattices. In: Paterson, K.G. (ed.) EUROCRYPT 2011. LNCS, vol. 6632, pp. 27–47. Springer, Heidelberg (2011)
15. Wang, J., Bi, J.: Lattice-based Identity-Based Broadcast Encryption, Cryptology ePrint Archive, Report 2010/288 (2010)

Supporting Secure Provenance Update by Keeping "Provenance" of the Provenance

Amril Syalim[1], Takashi Nishide[2], and Kouichi Sakurai[2]

[1] Fakultas Ilmu Komputer, Universitas Indonesia
amril.syalim@cs.ui.ac.id
[2] Department of Informatics, Kyushu University, Fukuoka, Japan
{nishide,sakurai}@inf.kyushu-u.ac.jp

Abstract. Provenance of data is a documentation of the origin and processes that produce the data. Many researchers argue that the provenance should be immutable: once a provenance is submitted, it should not be changed or updated. A main reason is that the provenance represents the history of data, and the history should not be altered or changed because it represents the fact in the past. Provenance can be represented by a graph, where each node represents the process executed by a party and an edge represents the relationship between two nodes (i.e. a child node uses the outputs of the parent nodes). A method to ensure that the provenance has not been updated is by using signature chain, where the signatures of the parent nodes are recorded in the children nodes so that any changes to the parent nodes will raise inconsistencies between the parent and the children. However, sticking to the requirement that the provenance should be immutable requires unlimited data storage and also we have problems whenever we need to update the provenance for an accidental error. In this paper, we propose a method that allows updates in the signature chain-based secure provenance, while keeping the signature consistent. The main idea is by keeping the "provenance" of the provenance itself, that is the history of update of the provenance, in the form of the signatures of the previous versions of the nodes. We implement the idea by keeping the signatures of the previous version in a signature tree similar to the Merkle-tree, where the a parent node in tree is the aggregate signature of the children. Using this method, the storage requirement to store signatures is always smaller than the number of updates.

Keywords: Provenance, Provenance Security, E-science, Data lineage.

1 Introduction

Provenance is a documentation of data history. It records the processes that produce the data and relationship between the processes. A simple example of the provenance is the documentation about how to produce a patient record in a hospital [2,6,12,13]. The patient record contains the data about the medical treatment or medical test of the patient which is produced by the physicians and laboratory staffs in the hospital. In this case the processes to produce data are executed by the physicians or the laboratory staffs. A process may have relationships with the other processes, for example: to produce a medical treatment, a

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 363–372, 2013.

physician uses the data produced by another physicians (i.e. medical diagnosis) or data used by a laboratory staffs (i.e. a result of blood or urine test of the patient).

Provenance can be represented by a graph [14,7,1,5], where the nodes represent the processes and the data while the edges represent relationships between processes. An example of the graph model for provenance is the open provenance model which is proposed as a standard for provenance [16,15]. The open provenance model specification defines some types of the nodes (i.e. actor, process and artifacts), and some types of the relationships between the nodes.

Some of the the important research problems with the provenance are related to security [21,10,11,12,19,20]. The security problems are caused by malicious update and deletion to the provenance. For example in a hospital, if we allow update to the provenance and the patient record, there can be inconsistencies between provenance and data created by a physician with another physician or laboratory staffs. A physician A may decide a medical treatment based on the diagnosis created by physician B, or medical test produced by a laboratory staff C. If B or C updates/deletes the provenance of diagnosis or medical test, the treatment by A may be incorrect because it is based on the previous versions of the diagnosis or medical tests produced by B or C. If the patient who takes the medical treatment complains about a misconduct by A, A cannot refer to B or C to explain why he/she decides the treatment. In an extreme case, it is possible B or C may maliciously/intentionally try to frame A for a misconduct charge.

1.1 Problem Definition

As a history record, many researchers argue that the provenance should be immutable (no change is allowed in the history records) [5,18,15,16,8,7,17]. However, without being able to update a submitted provenance, the provenance storage is always growing (indefinitely). In this paper, we try to address the problem on how to allow updates/deletes the provenance to save spaces on the storages, while keeping the provenance consistent. The methods are useful in the case the actors who produce the provenance honestly update the provenance for an accidental/unintentional errors and in the case the storage is limited. The main security requirement is: no actor is allowed to exploit the update/deletion features to do a malicious behaviors or avoid responsibilities for the data produced in the past.

1.2 Contributions

The idea of our solution is by keeping the "provenance" of the provenance, where using this method, we allow updates/deletions to the provenance but we keep the history of the previous versions of the provenance in the forms of the signatures of the previous versions. So that, even if an actor has updated the provenance and the data, the actor cannot reject the previous versions of the provenance and data. To save the spaces, we store the signatures in a tree similar to the Merkle-tree and combines the signatures using the aggregate signatures techniques. Using this method, for any number of updates the grow of signatures

storage for the previous versions of the provenance and data is much smaller than the normal updates in the provenance (where we should keep all signatures and data of the previous versions).

2 Related Work

2.1 Hash/Signature Chain, Stamping and Countering

The integrity scheme to timestamp digital documents using hash/signature chain was first proposed by Haber el al. [9] . It uses a Trusted Timestamping Service (TTS) that issues signed timestamps and also links two timestamps requested consecutively. The TSS links two timestamps by storing the hash value of the first timestamp in the second timestamp. Any changes to the first timestamp can be detected by checking the hash value in the second timestamp. This method is applied by Hasan et al. [12] to a chain model of the provenance where the provenance is modeled as a chain. Aldeco-Perez et al. [1] and Syalim et al. [20,19] extend the hash/signature chain to the provenance graph model.

To detect the problem of the deletion to a provenance node, Syalim et al. proposed countering provenance [20]. The basic idea is a Trusted Counter Server (TCS) assigns a unique consecutive counter number for each node in the provenance graph, so that any deletion to the nodes can be detected from the missing counter number in the nodes. To detect deletion of the newest node, the TCS should store the latest counter number in a trusted storage.

2.2 Aggregate Signatures

Aggregate signatures is a technique to combine signatures on many different messages into a short signature. Some aggregate signatures have restriction that they can only be verified if there is no duplicate messages or public keys. However, it is possible to develop a scheme that does not have any restriction [3].

Aggregate signatures can be implemented using bilinear maps [4], that is with the requirement of the existence of a mapping between groups for example the map $e : G_1 \times G_2 \rightarrow G_T$ where $|G_1| = |G_2| = |G_T|$ with bilinear (for all $u \in G_1, v \in G_2$ and $a, b \in \mathbb{Z}, e(u^a, v^b) = e(u, v)^{ab}$) and non-degenerate ($e(g_1, g_2) \neq 1$) properties. A particular aggregate signature scheme proposed by Boneh et al [4] is as follows:

Key Generation the user picks random secret key $x \overset{R}{\leftarrow} \mathbb{Z}_p$ and computes the public key $v \leftarrow g_2^x$
Signing to sign a message M, compute $h \leftarrow H(M)$, where $h \in G_1$, and the signature $\sigma \leftarrow h^x$.
Aggregation for a set of signatures $\{\sigma_1, \sigma_2, ..., \sigma_k\}$, compute the aggregate signature $\sigma \leftarrow \prod_{i=1}^{k} \sigma_i$.
Aggregate Verification for all users $u_i \in U$ with public keys $v_i \in G_1$ and the original messages M_i, computes $h_i \leftarrow H(M_i)$ and accept if $e(\sigma, g_2) = \prod_{i=1}^{k} e(h_i, v_i)$ holds.

3 Preliminaries

3.1 Definition

Definition 1 (Provenance definition). *Provenance related to the data set* $D = \{d_0, d_1, ..., d_{n-1}\}$ *stored in a database* DB *for* $n > 0$ *is a set of provenance nodes* $P = \{p_0, p_1, ..., p_{n-1}\}$ *and a binary relation* E *recorded in a Provenance Store* PS*. A provenance node* p_i *consists of an identification* PID*, a process description* A*, a process executor* CID*, the list of references to a set of inputs* $\{ref(I_i)\}$*, for* $I \subseteq D$ *and a reference to an output* $ref(d_i)$ *for* $d_i \in D$*.* E *is a binary relation in* $PIDSET$ *where* $PIDSET$ *is a set of* PID *of all provenance nodes.* E *represents the provenance edges such that for* $(x, y) \in PIDSET \times PIDSET$ *and for* $x \neq y$ *the process documented in provenance node with* PID y *takes the output of process documented in node with* PID x *as its input.*

In this definition, we store four kind of information about the process execution in the provenance: the description of the process, the information about the process executor, the list of the inputs and an output.

3.2 Participants in the Provenance System

The provenance system consists of the following participants:

- A data storage DB, where the data is stored
- A provenance storage PS, where the provenance of data is stored for a long term storage
- Process executors C with identification CID is the actors who execute the process, produce the data, submit data to DB, and submit the provenance to PS
- An auditor ADT, is the actor who checks the integrity of the provenance and data

3.3 The Basic Provenance Recording Method

A simple model of the provenance recording system is as follows. The process executor C_i queries the inputs $I \subseteq D$ from DB and a set of $\{PID_{in}\}$, that is the provenance ID of I. C_i executes the process that produces the data output d_i, creates the provenance node p_i and a set of edges, stores d_i to DB and submits provenance node p_i and edges to PS. The provenance edges are a set of mapping from the PID of provenance p_i to the parents that produce I, so we write the edges as $\{PID, PID_{in}\}$. We define the simple protocol as follows [Note: we use the index $_{in}$ to refer to the parents that produce the inputs I, for example PID_{in} is a PID of the parent, C_{in} is a process executor at the parent]:

$$
\begin{aligned}
&DB \rightarrow C_i : I \\
&PS \rightarrow C_i : \{PID_{in}\} \\
&C_i \rightarrow DB : d_i \\
&C_i \rightarrow PS : node = p_i, edge = \{PID, PID_{in}\}
\end{aligned}
$$

3.4 Aggregate Signature *Aggr* and *TreeAggr*

We define a function $Aggr$ which aggregates the signatures of different messages created by one or many different parties into one signature. A function $TreeAggr$ aggregates the signatures into a set of signatures where the number of element is less or equal to the total number of the original signatures. The method to aggregate the signatures using $TreeAggr$ is explained in Section 5.

4 Integrity Scheme for Provenance by Keeping "The Provenance" of the Provenance

4.1 Provenance Recording: Securing Provenance with Signature-Chain

For a secure provenance recording, when querying inputs I and the provenance of the inputs PID_{in}, the process executor also queries the collection of signature of inputs, that is the *signnode* in all parent nodes. In the scheme, we include a new participant, a Trusted Provenance Mediator (TPM), which is a trusted party who mediates the provenance recording. We assume that this party is trusted and will not cheat for any purposes. The complete provenance recording protocol is as follows:

$$
\begin{aligned}
DB \rightarrow C_i &: I \\
PS \rightarrow C_i &: \{PID_{in}, Sign_{C_{in}}(p_{in}, d_{in})\} \\
C_i \rightarrow DB &: d_i \\
C_i \rightarrow TPM &: Insert(node = p_i, edge = \{PID, PID_{in}\}, \\
&\quad signnode = Sign_{C_i}(p_i, d_i), \\
&\quad signedge = \{Sign_{C_{in}}(p_{in}, d_{in})\}) \\
TPM \rightarrow PS &: node = p_i, edge = \{PID, PID_{in}\}, \\
&\quad signnode = Sign_{C_i}(p_i, d_i), \\
&\quad signedge = \{Sign_{C_{in}}(p_{in}, d_{in})\}, \\
&\quad signtpm = Sign_{TPM}(signnode, signedge)
\end{aligned}
$$

The main difference of this scheme with the basic provenance recording method in Section 3.3 is we include signature chain to the provenance, by recording the signature of the node (*signnode*) and signatures of all edges that connect the node to the parents (*signedge* – represented by the parent signatures). The signatures are signed by the TPM before submitted to the provenance store PS (*signtpm*).

4.2 Provenance Update: Allowing Update in the Signature Chain

Update to a node by the same process executor is allowed, but we should keep the aggregate of the previous versions of the node. For example, C_i updates p_i

to p'_i and d_i to d'_i and no update to the edges, so that no change to *signedge*. This update will delete the previous version of p_i, and change it to p'_i. However it records the signatures of the previous versions in the form of an aggregate signature.

$$\begin{aligned}
C_i \rightarrow DB &: Update(d'_i) \\
C_i \rightarrow TPM &: Update(node = p'_i, edge = \{PID, PID_{in}\}, \\
&\quad signnode = Sign_{C_i}(p'_i, d'_i), \\
&\quad signedge = \{Sign_{C_{in}}(p_{in}, d_{in})\}) \\
TPM \rightarrow PS &: node = p'_i, edge = \{PID, PID_{in}\}, \\
&\quad signnode = Aggr\{Sign_{C_i}(p_i, d_i)\}, Sign_{C_i}(p'_i, d'_i)), \\
&\quad signedge = \{Sign_{C_{in}}(p_{in}, d_{in})\}, \\
&\quad signtpm = Sign_{TPM}(signnode, signedge)
\end{aligned}$$

The second update by C_i changes p'_i to p''_i and d'_i to d''_i and no change to *signedge*.

$$\begin{aligned}
C_i \rightarrow DB &: Update(d''_i) \\
C_i \rightarrow TPM &: Update(node = p''_i, edge = \{PID, PID_{in}\}, \\
&\quad signnode = Sign_{C_i}(p''_i, d''_i), \\
&\quad signedge = \{Sign_{C_{in}}(p_{in}, d_{in})\}) \\
TPM \rightarrow PS &: node = p''_i, edge = \{PID, PID_{in}\}, \\
&\quad signnode = Aggr\{Sign_{C_i}(p_i, d_i), Sign_{C_i}(p'_i, d'_i))\}, Sign_{C_i}(p''_i, d''_i)) \\
&\quad signedge = \{Sign_{C_{in}}(p_{in}, d_{in})\}, \\
&\quad signtpm = Sign_{TPM}(signnode, signedge)
\end{aligned}$$

We can accept any number of updates, however we only store the signatures of the previous versions of the nodes whose outputs are used by at least one child node.

4.3 The Case of Updates of the Parent Nodes

In Section 4.2, we assume no updates to *signedge* which means that no change to the input used by the node. However, a child may update the parent to the newest outputs of the parent. In this case, the child should keep the signature of the previous version of the parent using $TreeAggr$. For example, C_{in} updates p_{in} to p'_{in} and d_{in} to d'_{in}. C_{in} does exactly the same update as described in Section 4.2. After C_{in} updates the provenance and data, C_i also wants to update the parent to the newest version. So, C_i update *signedge* to $\{Sign_{C_{in}}(p'_{in}, d'_{in})\}$. The scheme to update the provenance in this case is as follows.

$$
\begin{aligned}
DB \rightarrow C_i &: I \\
PS \rightarrow C_i &: \{PID_{in}, Sign_{C_{in}}(p'_{in}, d'_{in})\} \\
C_i \rightarrow DB &: Update(d'_i) \\
C_i \rightarrow TPM &: Update(node = p'_i, edge = \{PID, PID_{in}\}, \\
&\quad signnode = Sign_{C_i}(p'_i, d'_i), \\
&\quad signedge = \{Sign_{C_{in}}(p'_{in}, d'_{in})\}) \\
TPM \rightarrow PS &: node = p'_i, edge = \{PID, PID_{in}\}, \\
&\quad signnode = Aggr\{Sign_{C_i}(p_i, d_i)\}, Sign_{C_i}(p'_i, d'_i)), \\
&\quad signedge = \{TreeAggr\{Sign_{C_{in}}(p_{in}, d_{in}), Sign_{C_{in}}(p'_{in}, d'_{in})\}\}, \\
&\quad signtpm = Sign_{TPM}(signnode, signedge)
\end{aligned}
$$

In this case, we aggregate all versions of *signedge* using $TreeAggr$.

4.4 Signature Verification

To verify consistency between a node and all of its children, the auditor *ADT* queries the *signnode* on the parent and all *signedge*(s) on the children. The *ADT* combines the *signedge*(s) to get an aggregate signatures of the parent and compares the result with *signnode*. A detailed example is described in Section 5.

5 Aggregating Signatures of Previous Versions of a Provenance Node Using *TreeAggr*

In the provenance, we use the aggregate signature to save the spaces needed for storing the signatures of all previous versions of the provenance nodes. Rather than verifying the aggregate signatures using the messages and the public key of the signers, we verify the signatures by comparing the aggregate signatures stored at a node with the aggregate of all signatures stored at all children of the nodes (we cannot verify the signatures by checking the messages – the previous versions of the provenance – because they have been deleted to save the spaces in each update). The motivation to use $TreeAggr$ is because each child may not store all versions of the parent. So, if we use normal aggregate signatures $Aggr$ to save the spaces, and stores the aggregates in each child, we may not be able to combine the signatures to form a full aggregate of all versions of the parent.

Using $TreeAggr$, at first we represent all versions of the parent in all leave nodes in the signature tree. The child can aggregate the signatures of two consecutive versions (even and odd) in the leave nodes and stores the aggregate in the parent of the leaves. The same method is applied for each level of nodes in the tree to combine the signatures into a smaller number of signatures stored in the nodes at the higher level until we cannot get two consecutive nodes to aggregate.

We give an example of the aggregation as follows. The node Q created by a process executor C has four versions: Q, Q', Q'', Q''', it has 6 children R, S, T, U, V, W. R and S use the output of the first, second and third versions of Q, while T and U only uses the output of the second version. V and W use the output of

the third version of Q. After sometimes, the children nodes S, U and W update the parent version to the latest version (Q''').

In the above case, in Q, we store the aggregate signatures of the previous version of Q, that is $Aggr\{Sign_C(Q), Sign_C(Q'), Sign_C(Q''\}$ and the latest signatures ($Sign_C(Q''')$). For each child, we store the $TreeAggr$ of all the parent signatures of the child. So, that for each child we store the signatures as follows:

$$\begin{aligned}
signedge_R &= TreeAggr\{Sign_C(Q), Sign_C(Q'), Sign_C(Q'')\} \\
&= Aggr\{Sign_C(Q), Sign_C(Q')\}, Sign_C(Q'') \\
signedge_S &= TreeAggr\{Sign_C(Q), Sign_C(Q'), Sign_C(Q''), Sign_C(Q''')\} \\
&= Aggr\{Sign_C(Q), Sign_C(Q')\}, Aggr\{Sign_C(Q''), Sign_C(Q''')\} \\
&= Aggr\{Sign_C(Q), Sign_C(Q'), Sign_C(Q''), Sign_C(Q''')\} \\
signedge_T &= TreeAggr\{Sign_C(Q')\} \\
&= Sign_C(Q') \\
signedge_U &= TreeAggr\{Sign_C(Q'), Sign_C(Q''')\} \\
&= Sign_C(Q'), Sign_C(Q''') \\
signedge_V &= TreeAggr\{Sign_C(Q'')\} \\
&= Sign_C(Q'') \\
signedge_W &= TreeAggr\{Sign_C(Q''), Sign_C(Q''')\} \\
&= Aggr\{Sign_C(Q''), Sign_C(Q''')\}
\end{aligned}$$

To check the integrity of the child nodes, we aggregate the signatures in some child nodes and compare the aggregate with the aggregate signatures on the parent. The parent Q stores an aggregate of the previous and the latest versions in $signnode = Aggr\{Sign_C(Q), Sign_C(Q'), Sign_C(Q''\}, Sign_C(Q''')$. Because the aggregation can be performed incrementally we may aggregate the signatures into

$$\begin{aligned}
signnode &= Aggr\{Aggr\{Sign_C(Q), Sign_C(Q'), Sign_C(Q'')\}, Sign_C(Q''')\} \\
&= Aggr\{Sign_C(Q), Sign_C(Q'), Sign_C(Q''), Sign_C(Q''')\}
\end{aligned}$$

We can compare the aggregate signatures on children with *signnode* stored at the parent by finding correct combination of the tree-aggregate on the children to form the same aggregate signature as stored in *signode*. Some examples of the correct combinations are shown as follows (note: we get $Sign_C(Q''')$ from $signedge_U$):

$$\begin{aligned}
signnode &= Aggr\{signedge_R, Sign_C(Q''')\} \\
&= signedge_S
\end{aligned}$$

6 Security Analysis

To show that a child is a node that uses an output of a version of the parent, the auditor *ADT* should query all *signedge* of all children of the node, find a combination of signatures that can reproduce *signnode* and compare with

signnode on the parent. Using the scheme explained in the previous versions, it is always possible to find a combination of signatures in the children of a node which produce exactly the same signature as *signnode*.

Theorem 1. *The Auditor ADT can always find a combination of the signatures* ***singedge*** *on the children which produce* ***signnode*** *in the parent.*

Proof. The parent node stores the signatures of the latest version and the signatures of the previous versions of the node, except for the versions that do not have any children. So, that a signature of a previous version of the parent node is stored in at least one child node. Each child only aggregates the signature of two consecutive versions of the parent, so that there are only two cases of the signatures: (1) the signature is not combined with another signature, in this case the signature is stored in its original form (2) the signature is combined with another signature to form an aggregation of two signatures.

If a signature σ_i is not combined, we should find the other signature of the consecutive version in other children. Because the other signature should be stored in at least one other child, if the other signature is stored as the first case (no combination with other signature), we can combine with σ_i to get a combination of two consecutive signatures. If the other signature has been combined, it should have been combined with σ_i (the same signatures stored in the other child), so in all cases we can combine elements in all *singedge* in the children to form *signnode* at the parent. □

7 Storage Requirements

The storage for *signnode* is always constant for any number of updates to the node, which is two signatures for each node (one signature for the latest version, and a signature for aggregate of signatures of all previous versions). As of the storage for *signedge*, in the best case, a child aggregates all the signatures of all versions of the parent, so it only needs to store one signature for each parent. The worst case is if the child does not uses any two consecutive versions of the parent, so we cannot reduce the number of signatures in $TreeAggr$. In this case, the number of signatures is $n/2$ where n is the number of updates to parent.

References

1. Aldeco-Pérez, R., Moreau, L.: Securing Provenance-Based Audits. In: McGuinness, D.L., Michaelis, J.R., Moreau, L. (eds.) IPAW 2010. LNCS, vol. 6378, pp. 148–164. Springer, Heidelberg (2010)
2. Álvarez, S., Vázquez-Salceda, J., Kifor, T., Varga, L.Z., Willmott, S.: Applying Provenance in Distributed Organ Transplant Management. In: Moreau, L., Foster, I. (eds.) IPAW 2006. LNCS, vol. 4145, pp. 28–36. Springer, Heidelberg (2006)
3. Bellare, M., Namprempre, C., Neven, G.: Unrestricted Aggregate Signatures. In: Arge, L., Cachin, C., Jurdziński, T., Tarlecki, A. (eds.) ICALP 2007. LNCS, vol. 4596, pp. 411–422. Springer, Heidelberg (2007)
4. Boneh, D., Gentry, C., Lynn, B., Shacham, H.: Aggregate and Verifiably Encrypted Signatures from Bilinear Maps. In: Biham, E. (ed.) EUROCRYPT 2003. LNCS, vol. 2656, pp. 416–432. Springer, Heidelberg (2003)

5. Braun, U., Shinnar, A., Seltzer, M.: Securing provenance. In: The 3rd USENIX Workshop on Hot Topics in Security (HOTSEC 2008), USENIX HotSec, pp. 1–5. USENIX Association, Berkeley (2008)
6. Deora, V., Contes, A., Rana, O.F., Rajbhandari, S., Wootten, I., Kifor, T., Varga, L.Z.: Navigating provenance information for distributed healthcare management. In: Web Intelligence, pp. 859–865 (2006)
7. Groth, P.T.: The Origin of Data: Enabling the Determination of Provenance in Multi-institutional Scientific Systems through the Documentation of Processes. PhD thesis, University of Southampton (2007)
8. Groth, P.T., Moreau, L.: Recording process documentation for provenance. IEEE Trans. Parallel Distrib. Syst. 20(9), 1246–1259 (2009)
9. Haber, S., Stornetta, W.S.: How to time-stamp a digital document. J. Cryptology 3(2), 99–111 (1991)
10. Hasan, R., Sion, R., Winslett, M.: Introducing secure provenance: problems and challenges. In: ACM Workshop on Storage Security and Survivability (StorageSS 2007), pp. 13–18 (2007)
11. Hasan, R., Sion, R., Winslett, M.: The case of the fake picasso: Preventing history forgery with secure provenance. In: 7th Conference on File and Storage Technologies (FAST 2009), pp. 1–14 (2009)
12. Hasan, R., Sion, R., Winslett, M.: Preventing history forgery with secure provenance. ACM Transactions on Storage 5(4), 12:1–12:43 (2009)
13. Kifor, T., Varga, L.Z., Vázquez-Salceda, J., Álvarez-Napagao, S., Willmott, S., Miles, S., Moreau, L.: Provenance in agent-mediated healthcare systems. IEEE Intelligent Systems 21(6), 38–46 (2006)
14. Miles, S., Groth, P.T., Munroe, S., Jiang, S., Assandri, T., Moreau, L.: Extracting causal graphs from an open provenance data model. Concurrency and Computation: Practice and Experience 20(5), 577–586 (2008)
15. Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P.T., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E.G., Van den Bussche, J.: The open provenance model core specification (v1.1). Future Generation Comp. Syst. 27(6), 743–756 (2011)
16. Moreau, L., Freire, J., Futrelle, J., McGrath, R.E., Myers, J., Paulson, P.: The Open Provenance Model: An Overview. In: Freire, J., Koop, D., Moreau, L. (eds.) IPAW 2008. LNCS, vol. 5272, pp. 323–326. Springer, Heidelberg (2008)
17. Muniswamy-Reddy, K.-K.: Foundations for Provenance-Aware Systems. PhD thesis, Harvard University (2010)
18. Simmhan, Y., Plale, B., Gannon, D.: A survey of data provenance in e-science. In: SIGMOD Record, pp. 31–36 (2005)
19. Syalim, A., Nishide, T., Sakurai, K.: Preserving Integrity and Confidentiality of a Directed Acyclic Graph Model of Provenance. In: Foresti, S., Jajodia, S. (eds.) Data and Applications Security and Privacy XXIV. LNCS, vol. 6166, pp. 311–318. Springer, Heidelberg (2010)
20. Syalim, A., Nishide, T., Sakurai, K.: Securing provenance of distributed processes in an untrusted environment. IEICE Transactions 95-D(7), 1894–1907 (2012)
21. Tan, V., Groth, P.T., Miles, S., Jiang, S., Munroe, S.J., Tsasakou, S., Moreau, L.: Security Issues in a SOA-Based Provenance System. In: Moreau, L., Foster, I. (eds.) IPAW 2006. LNCS, vol. 4145, pp. 203–211. Springer, Heidelberg (2006)

New Ciphertext-Policy Attribute-Based Access Control with Efficient Revocation

Xingxing Xie[1], Hua Ma[1], Jin Li[2], and Xiaofeng Chen[3,⋆]

[1] Department of Mathematics,
Xidian University, Xi'an 710071, P.R. China
xiexingxing11@163.com, mahua@126.com
[2] Department of Computer Science,
Guangzhou University, Guangzhou 510006, P.R. China
jinli71@gmail.com
[3] State Key Laboratory of Integrated Service Networks (ISN),
Xidian University, Xi'an 710071, P.R. China
xfchen@xidian.edu.cn

Abstract. Attribute-Based Encryption (ABE) is one of the promising cryptographic primitives for fine-grained access control of shared outsourced data in cloud computing. However, before ABE can be deployed in data outsourcing systems, it has to provide efficient enforcement of authorization policies and policy updates. However, in order to tackle this issue, efficient and secure attribute and user revocation should be supported in original ABE scheme, which is still a challenge in existing work. In this paper, we propose a new ciphertext-policy ABE (CP-ABE) construction with efficient attribute and user revocation. Besides, an efficient access control mechanism is given based on the CP-ABE construction with an outsourcing computation service provider.

Keywords: Attribute-based encryption, revocation, outsourcing, re-encryption.

1 Introduction

As a relatively new encryption technology, attribute-based encryption(ABE) has attracted lots of attention because ABE enables efficient one-to-many broadcast encryption and fine-grained access control system. Access control is one of the most common and versatile mechanisms used for information systems security enforcement. An access control model formally describes how to decide whether an access request should be permitted or repudiated. Particularly, in the outsourcing environment, designing an access control will introduce many challenges.

However, the user and attribute revocation is still a challenge in existing ABE schemes. Many schemes [1,3,9] are proposed to cope with attribute-based access control with efficient revocation. The most remarkable is the scheme proposed by

⋆ Corresponding author.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 373–382, 2013.

J.Hur and D.K.Noh, which realizes attribute-based access control with efficient fine-grained revocation in outsourcing. However, in the phase of key update, the data service manager will perform heavy computation at every time of update, which could be a bottleneck for the data service manager. Moreover, in the outsourcing environment, external service provider [14,15] is indispensable. Thus, in this paper, we attempt to solve the problem of efficient revocation in attribute-based data access control using CP-ABE for outsourced data.

1.1 Related Work

For the ABE, key-policy ABE (KP-ABE) and ciphertext-policy ABE (CP-ABE) are more prevalent than the others. To take control of users' access right by a data owner, we specify CP-ABE as the data outsourcing architecture.

Attribute Revocation. Recently, several attribute revocable ABE schemes have been announced [2,3,4]. Undoubtedly, these approaches have two main problems, which consists of security degradation in terms of the backward and forward security [1,5]. In the previous schemes, the key authority periodically announce a key update, that will lead to a bottleneck for the key authority. Two CP-ABE schemes with immediate attribute revocation with the help of semi-honest service provider are proposed in [6,7]. However, achieving fine-grained user access control failed. Junbeom et al. [1] proposed a CP-ABE scheme with fine-grained attribute revocation with the help of the honest-but-curious proxy deployed in the data service provider.

User Revocation. In [8], a fine-grained user-level revocation is proposed using ABE that supports negative clause. In the previous schemes [8,9], a user loses all the access rights to the data when he is revoked from a single attribute group. Attrapadung and Imai [10] suggested another user-revocable ABE schemes, in which the data owner should take full control of all the membership lists that leads to be not applied in the outsourcing environments.

1.2 Our Contribution

In this study, aiming at reducing the overhead computation at data service manager, we propose an ciphertext policy attribute-based access control with efficient revocation. This construction is based on a CP-ABE construction with efficient user and attribute revocation. Compared with [1], in our proposed construction, in the phase of key update, the computation operated by the data service manager will reduce by half. In Table 1 we summarize the comparisons between our proposed scheme and [1] in terms of the computations in the phase of key update. Furthermore, we formally show the security proof based on security requirement in the access control system.

2 Systems and Models

2.1 System Description and Assumptions

There are four entities involved in our attribute-based access control system:

- Trusted authority. It is the party that is fully trusted by all entities participating in the data outsourcing system.
- Data owner. It is a client who owns data and encrypts the outsourced data.
- User. It is an entity who would like to access the cryptographic data.
- Service provider. It is an entity that provides data outsourcing service. The data servers are responsible for storing the outsourced data. Access control from outside users is executed by the data service manager, which is assumed to be honest-but-curious.

2.2 Threat Model and Security Requirements

- Data confidentiality. It is not allowed to access the plaintext if a user's attributes do not satisfy the access policy. In addition, unauthorized data service manager should be prevented from accessing the plaintext of the encrypted data that it stores.
- Collusion-resistance. Even if multiple users collaborate, they are unable to decrypt encrypted data by combining their attribute keys.
- Backward and forward secrecy. In our context, backward secrecy means that if a new key is distributed for the group when a new member joins, he is not able to decrypt previous messages before the new member holds the attribute. On the other hand, forward secrecy means that a revoked or expelled group member will not be able to continue accessing the plaintext of the subsequent data exchanged (if it keeps receiving the messages), when the other valid attributes that he holds do not satisfy the access policy.

3 Preliminaries and Definitions

3.1 Bilinear Pairings

Let $\mathbb{G}$ and $\mathbb{G}_T$ be two cyclic group of prime order p. The Discrete Logarithm Problem on both $\mathbb{G}$ and $\mathbb{G}_T$ are hard. A bilinear map e is a map function e: $\mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$ with the following properties:

1. Bilinearity: For all $A, B \in \mathbb{G}$, and $a, b \in \mathbb{Z}_p^*$, $e(A^a, B^b) = e(A, B)^{ab}$.
2. Non-degeneracy: $e(g, g) \neq 1$, where g is the generator of $\mathbb{G}$.
3. Computability: There exsits an efficient algorithm to compute the pairing.

3.2 Decisional Bilinear Diffie-Hellman Exponent (BDHE) Assumption [12]

The decisional BDHE problem is to compute $e(g,g)^{a^{q+1}s} \in \mathbb{G}_T$, given a generator g of $\mathbb{G}$ and elements $\overrightarrow{y} = (g_1, \cdots, g_q, g_{q+1}, \cdots, g_{2q}, g^s)$ for $a, s \in \mathbb{Z}_p^*$. Let g_i denote g^{a^i}.

An algorithm $\mathcal{A}$ has advantage $\epsilon(\kappa)$ in solving the decisional BDHE problem for a bilinear map group $\langle p, \mathbb{G}, \mathbb{G}_\mathbb{T}, e\rangle$, where κ is the security parameter, if $|Pr[\mathcal{B}(\overrightarrow{y}, g, T = e(g,g)^{a^{q+1}s}) = 0] - Pr[\mathcal{B}(\overrightarrow{y}, T = R) = 0]| \geq \epsilon(\kappa)$.

$\langle p, \mathbb{G}, \mathbb{G}_\mathbb{T}, e\rangle$ is deemed to satisfy the decisional BDHE assumption. when for every polynomial-time algorithm (in the security parameter κ) to solve the decisional BDHE problem on $\langle p, \mathbb{G}, \mathbb{G}_\mathbb{T}, e\rangle$, the advantage $\epsilon(\kappa)$ is a negligible function.

3.3 System Definition and Our Basic Construction

Let $\mathcal{U} = \{u_1, \cdots, u_n\}$ be the whole of users. Let $\mathcal{L} = \{1, \cdots, p\}$ be the universe of attributes that defines, classifies the user in the system. Let $G_i \subset \mathcal{U}$ be a set of users that hold the attribute i. Let $\mathcal{G} = \{G_1, \cdots, G_p\}$ be the whole of such attribute groups. Let K_i be the attribute group key that is possessed by users, who own the attribute i.

Ciphertext Policy Attribute-Based Access Control with User Revocation.

Definition 1. *A CP-ABE with user revocation capability scheme consists of six algorithms:*

- **Setup:** *Taking a security parameter k, this algorithm outputs a public key PK and a master secret key MK.*
- **KeyGen(MK, S, U):** *Taking the MK, and a set of attributes $S \subseteq \mathcal{L}$ and users $U \subseteq \mathcal{U}$, this algorithm outputs a set of private attributes keys SK for each user.*
- **KEKGen(U):** *Taking a set of users $U \subseteq \mathcal{U}$, this algorithm outputs KEKs for each user, which will be used to encrypt attribute group keys.*
- **Encrypt(PK, M, $\mathcal{T}$):** *Taking the PK, a message M and an access structure $\mathcal{T}$, this algorithm outputs the ciphertext CT .*
- **Re-Encrypt(CT, G):** *Taking ciphertext CT and attributes groups G, this algorithm outputs re-encrypted CT'.*
- **Decrypt(CT', SK, K_S):** *The decryption algorithm takes as input the ciphertext $CT^{'}$, a private key SK, and a set of attribute group keys K_S. The decryption can be done.*

4 Ciphertext Policy Attribute-Based Access Control with Efficient Revocation

4.1 KEK Construction

In our scheme, KEK tree will be used to re-encrypt the ciphertext encrypted by the owner, which is constructed by the data service manager as in Fig.1. Now, some basic properties of the KEK tree will be presented as [1].

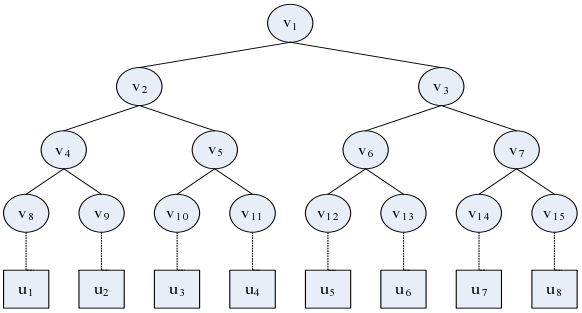


Fig. 1. KEK tree attribute group key distribution

4.2 Our Construction

Let $e : \mathbb{G} \times \mathbb{G} \longrightarrow \mathbb{G}_T$ be a bilinear map of prime order p with the generator g. A security parameter, κ, will decide the size of the groups. We will additionally employ a hash function $H : \{0,1\}^* \longrightarrow \mathbb{G}$ that we will model as a random oracle.

System Setup and Key Generation. The trusted authority (TA) first runs Setup algorithm by choosing a bilinear map : $\mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ of prime order δ with a generator g. Then, TA chooses two random $\alpha, a \in Z_p$. The public parameters are $PK = \{\mathbb{G}, g, h = g^a, e(g,g)^\alpha\}$. The master key is $MK = \{g^\alpha\}$, which is only known by the TA.

After executing the Setup algorithm producing PK and MK, each user in U needs to register with the TA, who verifies the user's attributes and issues proper private keys for the user. Running $KeyGen(MK, S, U)$, the TA inputs a set of users $U \subseteq \mathcal{U}$ and attributes $S \subseteq \mathcal{L}$, and outputs a set of private key components corresponding to each attribute $j \in S$. The key generation algorithm is presented as follows:

1. Choose a random $r \in \mathbb{Z}_p^*$, which is unique to each user.
2. Compute the following secret value to the user $u \in U$ as:

$$SK = (K = g^\alpha g^{ar}, L = g^r, \forall j \in S : D_j = H(j)^r)$$

After implementing above operations, TA sends the attribute groups G_j [1] for each $j \in S$ to the data service manager.

KEK Generation. After obtaining the attribute groups G_j for each $j \in S$ from the TA, the data service manager runs $KEKGen(U)$ and generates KEKs for users in U. Firstly, the data service manager sets a binary KEK tree for the unniverse of users $\mathcal{U}$ just like that described above. The KEK tree is responsible for distributing the attribute group keys to users in $U \subseteq \mathcal{U}$. For instance, u_3 stores $PK_3 = \{KEK_{10}, KEK_5, KEK_2, KEK_1\}$ as its path keys in Fig.2.

Then, in the data re-encryption phase, the data service manager will encrypt the attribute group keys by no means the path keys, i.e. KEKs. The method of

the key assignment is that keys are assigned randomly and independently from each other, which is information theoretic.

Data Encryption. To encrypt the data M, a data user needs to specify a policy tree $\mathcal{T}$ over the universe of attributes $\mathcal{L}$. Running $Encrypt(PK, M, \mathcal{T})$, the data M is enforced attribute-based access control. The policy tree $\mathcal{T}$ is defined as follows.

For each node x in the tree $\mathcal{T}$, the algorithm chooses a polynomial q_x, which is chosen in a top-down manner, starting from the root node R and its degree d_x is one less than the threshold value k_x of the node, that is, $d_x = k_x - 1$. For the root node R, it randomly chooses an $s \in \mathbb{Z}_p^*$ and sets $q_R(0) = s$. Except the root node R, it sets $q_x(0) = q_{p(x)}(index(x))$ and chooses d_x other points randomly to completely define q_x for any other node x. Let Y be the set of leaf nodes in $\mathcal{T}$. The ciphertext is then constructed by giving the policy tree $\mathcal{T}$ and computing

$$CT = (\mathcal{T}, \widetilde{C} = Me(g,g)^{\alpha s}, C = g^s,$$

$$\forall y \in Y : C_y = g^{aq_x(0)} \cdot H(y)^{-s})$$

After constructing CT, the data owner outsources it to the service provider securely.

Data Re-Encryption. On receiving the ciphertext CT, the data service manager re-encrypts CT using a set of the membership information for each attribute group $G \subseteq \mathcal{G}$. The re-encryption algorithm progresses as follows:

1. For all $G_y \in G$, chooses a random $K_y \in \mathbb{Z}_p^*$ and re-encrypts CT as follows:

$$CT' = (\mathcal{T}, \widetilde{C} = Me(g,g)^{\alpha s}, C = g^s,$$

$$\forall y \in Y : C_y = (g^{aq_x(0)} \cdot H(y)^{-s})^{K_y})$$

2. After re-encrypting CT, the data service manager needs to employ a method to deliver the attribute group keys to valid users. The method we used is a symmetric encryption of a message M under a key K, in other words, $E_K : \{0,1\}^k \longrightarrow \{0,1\}^k$, as follow:

$$Hdr = (\forall y \in Y : \{E_K(K_y)\}_{K \in KEK(G_y)})$$

After above all operations, the data service manager responds with (Hdr, CT') to the user sending any data request.

Data Decryption. Data decryption phase consists of the attribute group key decryption from Hdr and message decryption.

Attribute Group Key Decrypt. To execute data decryption, a user u_t first decrypt the attribute group key for all attributes in S that the user holds from

Hdr. If the user $u_t \in G_j$, he can decrypt the attribute group key K_j from Hdr using a KEK that is possessed by the user. For example, if $G_j = \{u_1, u_2, u_5, u_6\}$ in Fig.2, u_5 can decrypt the K_j using the path key $KEK_6 \in PK_5$. Next, u_t updates its secret key as follows:

$$SK = (K = g^{\alpha} g^{ar}, \forall j \in S : D_j = H(j)^r, L = (g^r)^{1/K_j})$$

Message Decrypt. Once the user updates its secret key, he then runs the $Decrypt(CT', SK, K_S)$ algorithm as follows. The user runs a recursive function $DecryptNode(CT', SK, R)$, R is the root of $\mathcal{T}$. The recursion function is the same as defined in [2]. And if x is a leaf node, then $DecryptNode(CT', SK, x)$ is proceeded as follow when $x \in S$ and $u_t \in G_x$:

$$\begin{aligned} Decrypt(CT', SK, x) &= e(C_x, L) \cdot e(C, D_x) \\ &= e((H(x)^{-s} \cdot g^{aq_x(0)})^{K_x}, (g^r)^{1/K_x}) \cdot e(g^s, H(x)^r) \\ &= e(g, g)^{raq_x(0)} \end{aligned}$$

Now we consider the recursion when x is a nonleaf node processed as follows: $\forall z$ is the child of x, it calls $DecryptNode(CT', SK, z)$ and stores the output as F_z. Let S_x be an arbitrary k_x-sized set of child nodes z, then computes:

$$\begin{aligned} F_x &= \prod_{z \in S_x} F_z^{\Delta_{i,S'_x}(0)}, where^{i=index(x),}_{S'_x=\{index(z):z\in S_x\}} \\ &= \prod_{z \in S_x} (e(g, g)^{r \cdot aq_z(0)})^{\Delta_{i,S'_x}(0)} \\ &= \prod_{z \in S_x} (e(g, g)^{r \cdot aq_{p(z)}(index(z))})^{\Delta_{i,S'_x}(0)} \\ &= \prod_{z \in S_x} (e(g, g)^{r \cdot aq_x(i)})^{\Delta_{i,S'_x}(0)} \\ &= e(g, g)^{r \cdot aq_x(0)} \end{aligned}$$

Where $i = index(z)$ and $S'_x = \{index(z) : z \in S_x\}$. Finally, if x is the root node R of the access tree $\mathcal{T}$, the recursive algorithm returns $A = DecryptNode(CT', SK, R) = e(g, g)^{ras}$. And the algorithm decrypts the ciphertext by computing

$$\widetilde{C}/(e(C, K)/A) = \widetilde{C}/(e(g^s, g^{\alpha} g^{ra})/e(g, g)^{ras}) = M.$$

5 Key Update

In this section, when a user wants to leave or join several attribute groups, he needs to send the attributes changed to TA. Without loss of generality, assume there is any membership change in G_j, which is equal to that a user comes to hold or drop an attribute j at the some instance. Next, we progress the update procedure as follows:

1. The data service manager selects a random $s' \in \mathbb{Z}_p^*$ and a K_i', and re-encrypts the ciphertext CT' using PK as

$$\begin{aligned} CT^{'} &= (\mathcal{T}, \widetilde{C} = Me(g,g)^{\alpha(s+s')}, C = g^{(s+s')}, \\ C_i &= (g^{a(q_x(0)+s')} \cdot H(i)^{-s})^{K_i'} \\ &\quad \forall y \in Y \setminus \{i\} : C_y = (g^{a(q_x(0)+s')} \cdot H(y)^{-s})^{K_y}). \end{aligned}$$

2. After updating the ciphertext, the data service manager selects new minimum cover sets for G_i changed and generates a new header message as follows:

$$\begin{aligned} Hdr = &(\{E_K(K_i')\}_{K \in KEK(G_i)}, \\ &\forall y \in Y \setminus \{i\} : \{E_K(K_y)\}_{K \in KEK(G_y)}). \end{aligned}$$

6 Efficiency Analysis

In this section, we analyze the efficiency of the proposed schemes with the scheme [1]. Table 1 shows the comparisons between our scheme and scheme [1] in terms of the computations in the phase of key update. In our scheme, the number of exponentiations is reduced to ω+3. However, in the scheme [1], the number of exponentiations unexpectedly is 2ω+3. Thus, it will enormously improve computational efficiency.

Table 1. Result Comparison

	the Number of Exponentiations of Key Update
Scheme one	$2\omega + 3$
Our scheme	$\omega + 3$

7 Security

In this section, the security of the proposed scheme is given based on the security requirements discussed in Section 2.

Theorem 1. ***Collusion Resistance.*** *The proposed scheme is secure to resist collusion attack.*

Proof. In CP-ABE, the secret s sharing is embedded into a ciphertext, and to decrypt a ciphertext, a user or a colluding attacker needs to recover $e(g,g)^{\alpha s}$. To recover $e(g,g)^{\alpha s}$, the attacker must pair C_x from the ciphertext and D_x from the other colluding users private key for an attribute x. However, every user's private key is uniquely generated by a random r. Thus, even if the colluding users are all valid, the attacker can not recover $e(g,g)^{\alpha s}$.

Theorem 2. ***Data Confidentiality.*** *The proposed scheme prevents unauthorized users and the curious service provider from acquiring the privacy of the outsourced data.*

Proof. Firstly, if the attributes held by a user don't satisfy the tree policy $\mathcal{T}$, the user will not recover the value $e(g,g)^{ras}$, which leads the ciphertext not to be deciphered. Secondly, when a user is revoked from some attribute groups that satisfy the access policy, he will lose the updated attribute group key. If the user would like to decrypt a node x for corresponding attribute, he needs to pair C_x from the ciphertext and L encrypted by K_x from its private key. As the user cannot obtain the updated attribute group key K_x, he cannot decrypt the value $e(g,g)^{raq_x(0)}$. Ultimately, Since we assume that the service provider is honest-but-curious, the service provider cannot be totally trusted by users. The service provider is available to the ciphertext and each attribute group key. However, any of the private keys for the set of attributes are not given to the data service manager. Thus, the service provider will not decrypt the ciphertext.

Theorem 3. ***Backward and Forward Secrecy.*** *For backward and forward secrecy of the outsourced data, the proposed scheme is secure against any newly joining and revoked users, respectively.*

Proof. When a user comes to join some attribute groups, the corresponding attribute group keys are updated and delivered to the user securely. Even if the user has stored the previous ciphertext exchanged and the holding attributes satisfy the access policy, he cannot decrypt the pervious ciphertext. That is because, though he could succeed in computing $e(g,g)^{ra(s+s')}$, it will not help to recover the value $e(g,g)^{\alpha s}$ from the updated ciphertext.

Furthermore, when a user comes to leave some attribute groups, the corresponding attribute group keys are updated and not delivered to the user. As the user cannot obtain the updated attribute group keys, he cannot decrypt any nodes corresponding to the updated attributes. Moreover, even if the user has stored $e(g,g)^{\alpha s}$, he cannot decrypt the subsequent value $e(g,g)^{\alpha(s+s')}$. Because he is not available to random s'.

8 Conclusion

In this paper, aiming at improving the efficiency of revocation to make CP-ABE widely deployed in access control, we introduced a new CP-ABE construction. In this construction, the overall computation of key update become less. Furthermore, the security proof is also shown based on access control security requirements.

Acknowledgements. We are grateful to the anonymous referees for their invaluable suggestions. This work is supported by the National Natural Science Foundation of China (Nos. 60970144, 61272455 and 61100224), and China 111 Project (No. B08038).

References

1. Hur, J., Noh, D.K.: Attribute-based Access Control with Efficient Revocation in Data Outsourcing System. IEEE Transactions on Parallel and Distributed System, 1214–1221 (2011)
2. Bethencourt, J., Sahai, A., Waters, B.: Ciphertext-Policy Attribute-Based Encryption. In: Proc. IEEE Symp. Security and Privacy, pp. 321–334 (2007)
3. Boldyreva, A., Goyal, V., Kumar, V.: Identity-Based Encryption with Efficient Revocation. In: Proc. ACM Conf. Computer and Comm. Security, pp. 417–426 (2008)
4. Pirretti, M., Traynor, P., McDaniel, P., Waters, B.: Secure Attribute-Based Systems. In: Proc. ACM Conf. Computer and Comm. Security (2006)
5. Rafaeli, S., Hutchison, D.: A Survey of Key Management for Secure Group Communication. ACM Computing Surveys 35(3), 309–329 (2003)
6. Ibraimi, L., Petkovic, M., Nikova, S., Hartel, P., Jonker, W.: Mediated Ciphertext-Policy Attribute-Based Encryption and Its Application. In: Proc. Int'l Workshop Information Security Architecture, CSAW 2007 (2007)
7. Yu, S., Wang, C., Ren, K., Lou, W.: Attribute Based Data Sharing with Attribute Revocation. In: Proc. ACM Symp. Information, Computer and Comm. Security, ASIACCS 2010 (2010)
8. Ostrovsky, R., Sahai, A., Waters, B.: Attribute-Based Encryption with Non-Monotonoc Access Structures. In: Proc. ACM Conf. Computer and Comm. Security, pp. 195–203 (2007)
9. Liang, X., Lu, R., Lin, X., Shen, X.: Ciphertext Policy Attribute Based Encryption with Efficient Revocation. Technical Report, Univ. of Waterloo (2011), `http://bbcr.uwaterloo.ca/x27liang/papers/abe/%20with%20revocation.pdf`
10. Attrapadung, N., Imai, H.: Conjunctive Broadcast and Attribute-Based Encryption. In: Shacham, H., Waters, B. (eds.) Pairing 2009. LNCS, vol. 5671, pp. 248–265. Springer, Heidelberg (2009)
11. Zhou, Z., Huang, D.: Efficient and Secure Data Storage Operations for Mobile Cloud Computing. Cryptology ePrint Archive, Report: 2011/185 (2011)
12. Waters, B.: Ciphertext-Policy Attribute-Based Encryption: An Expressive, Efficient, and Provably Secure Realization. Communications of the ACM, 53–70 (2011)
13. Shamir, A.: How to Share a Secret. Computer Science, 612–613 (1979)
14. Wang, J., Ma, H., Tang, Q., Lin, J., Zhu, H., Ma, S., Chen, X.: A New Efficient Verifiable Fuzzy Keyword Search Scheme. Journal of Wireless Mobile Networks, Ubiquitous Computing and Dependable Applications 3(4), 61–71 (2012)
15. Zhao, Y., Chen, X., Ma, H., Tang, Q., Zhu, H.: A New Trapdoor-indistinguishable Public Key Encryption with Keyword Search. Journal of Wireless Mobile Networks, Ubiquitous Computing and Dependable Applications 3(1/2), 72–81 (2012)

Provably Secure and Subliminal-Free Variant of Schnorr Signature

Yinghui Zhang[1,⋆], Hui Li[1], Xiaoqing Li[1], and Hui Zhu[1,2]

[1] State Key Laboratory of Integrated Service Networks (ISN), Xidian University, Xi'an 710071, P.R. China
yhzhaang@163.com
[2] Network and Data Security Key Laboratory of Sichuan Province, Chengdu 611731, P.R. China

Abstract. Subliminal channels present a severe challenge to information security. Currently, subliminal channels still exist in Schnorr signature. In this paper, we propose a subliminal-free variant of Schnorr signature. In the proposed scheme, an *honest-but-curious* warden is introduced to help the signer to generate a signature on a given message, but it is disallowed to sign messages independently. Hence, the signing rights of the signer is guaranteed. In particular, our scheme can completely close the subliminal channels existing in the random session keys of Schnorr signature scheme under the intractability assumption of the discrete logarithm problem. Also, the proposed scheme is proved to be existentially unforgeable under the computational Diffie-Hellman assumption in the random oracle model.

Keywords: Digital signature, Information hiding; Subliminal channel, Subliminal-freeness, Provable security.

1 Introduction

The notion of subliminal channels was introduced by Simmons [1]. He proposed a prison model in which authenticated messages are transmitted between two prisoners and known to a warden. The term of "subliminal" means that the sender can hide a message in the authentication scheme, and the warden cannot detect or read the hidden message. Simmons discovered that a secret message can be hidden inside the authentication scheme and he called this "hidden" communication channel as the subliminal channel. The "hidden" information is known as subliminal information.

As a main part of information hiding techniques [2–6], subliminal channels have been widely studied and applied [7–12]. However, they also present a severe challenge to information security. To the best of our knowledge, subliminal channels still exist in Schnorr signature [13].

Our Contribution. In this paper, we propose a subliminal-free variant of Schnorr signature scheme, in which an *honest-but-curious* warden is introduced

⋆ Corresponding author

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 383–391, 2013.

to help the signer to generate a signature on a given message, but it is disallowed to sign messages independently. In addition, the signer cannot control outputs of the signature algorithm. To be specific, the sender has to cooperate with the warden to sign a given message. Particularly, our scheme is provably secure and can completely close the subliminal channels existing in the random session keys in Schnorr signature scheme.

Related Work. Plenty of researches have been done on both the construction of subliminal channels and the design of subliminal-free protocols [7–11, 14–17]. Since the introduction of subliminal channels, Simmons [18] also presented several narrow-band subliminal channels that do not require the receiver to share the sender's secret key. Subsequently, Simmons [15] proposed a broad-band subliminal channel that requires the receiver to share the sender's secret key. For the purpose of information security, Simmons then proposed a protocol [19] to close the subliminal channels in the DSA digital signature scheme. However, Desmedt [14] showed that the subliminal channels in the DSA signature scheme cannot be completely closed using the protocol in [19]. Accordingly, Simmons adopted the cut-and-choose method to reduce the capacity of the subliminal channels in the DSA digital signature algorithm [20]. However, the complete subliminal-freeness still has not been realized. To be specific, the computation and communication costs significantly increase with the reduction of the subliminal capacity. On the other hand, subliminal channels in the NTRU cryptosystem and the corresponding subliminal-free methods [21] were proposed. Also, a subliminal channel based on the elliptic curve cryptosystem was constructed [8, 17]. As far as the authors know, the latest research is mainly concentrated on the construction [10, 11, 16] of subliminal channels and their applications [7, 12, 22, 23].

Outline of the Paper. The rest of this paper is organized as follows. In Section 2, we introduce some notations and complexity assumptions, and then discuss subliminal channels in probabilistic digital signature. In Section 3, we lay out the abstract subliminal-free signature specification and give the formal security model. The proposed provably secure and subliminal-free variant of Schnorr signature scheme is described in Section 4. Some security considerations are discussed in Section 5. Finally, we concludes the work in Section 6.

2 Preliminaries

2.1 Notations

Throughout this paper, we use the notations, listed in Table 1, to present our construction.

2.2 Complexity Assumptions

Discrete Logarithm Problem (DLP): Let $\mathbb{G}$ be a group, given two elements g and h, to find an integer x, such that $h = g^x$ whenever such an integer exists.

Table 1. Meaning of notations in the proposed scheme

Notation	Meaning
$s \in_R \mathbb{S}$	s is an element randomly chosen from a set $\mathbb{S}$.
l_s	the bit length of the binary representation of s.
$s_1 \parallel s_2$	the concatenation of bit strings s_1 and s_2.
$\gcd(a, b)$	the greatest common divisor of two integers a and b.
x^{-1}	the modular inverse of x modulo q such that $x^{-1}x = 1 \mod q$, where x and q are relatively prime, *i.e.*, $\gcd(x, q) = 1$.
$\mathbb{G}_{g,p}$	a cyclic group with order q and a generator g, where q is a large prime factor of $p-1$ and p is a large prime. That is, $\mathbb{G}_{g,p} = \{g^0, g^1, \cdots, g^{q-1}\} =< g >$, which is a subgroup in the multiplicative group $GF^*(p)$ of the finite field $GF(p)$.

Intractability Assumption of DLP: In group $\mathbb{G}$, it is computationally infeasible to determine x from g and h.
Computation Diffie-Hellman (CDH) Problem: Given a 3-tuple $(g, g^a, g^b) \in \mathbb{G}^3$, compute $g^{ab} \in \mathbb{G}$. An algorithm $\mathcal{A}$ is said to have advantage ϵ in solving the CDH problem in $\mathbb{G}$ if

$$\Pr\left[\mathcal{A}(g, g^a, g^b) = g^{ab}\right] \geq \epsilon,$$

where the probability is over the random choice of g in $\mathbb{G}$, the random choice of a, b in $\mathbb{Z}_q^*$, and the random bits used by $\mathcal{A}$.

CDH Assumption: We say that the (t, ϵ)-CDH assumption holds in $\mathbb{G}$ if no t-time algorithm has advantage at least ϵ in solving the CDH problem in $\mathbb{G}$.

2.3 Subliminal Channels in Probabilistic Digital Signature

Probabilistic digital signature [24] can serve as the host of subliminal channels. In fact, the subliminal sender can embed some information into a subliminal channel by controlling the generation of the session keys. After verifying a given signature, the subliminal receiver uses an extraction algorithm to extract the embedded information. Note that the extraction algorithm is only possessed by the authorized subliminal receiver. Hence, anyone else cannot learn whether there exists subliminal information in the signature [25], not to mention extraction of the embedded information.

In a probabilistic digital signature scheme, the session key can be chosen randomly, and hence one message may correspond to several signatures. More specifically, if different session keys are used to sign the same message, different digital signatures can be generated. This means that redundant information exists in probabilistic digital signature schemes, which creates a condition for subliminal channels. The subliminal receiver can use these different digital signatures to obtain the subliminal information whose existence can hardly be learnt by the others.

In particular, there exists subliminal channels in a typical probabilistic digital signature, namely Schnorr Signature [13].

3 Definition and Security Model

3.1 Specification of Subliminal-Free Signature

A subliminal-free signature scheme consists of three polynomial-time algorithms **Setup**, **KeyGen**, an interactive protocol **Subliminal-Free Sign**, and **Verify** below. Based on a subliminal-free signature scheme, a sender A performs an interactive protocol with a warden W. And, W generates the final signature σ and transmits it to a receiver B. Note that W is *honest-but-curious*. That is, W will honestly execute the tasks assigned by the related algorithm. However, it would like to learn secret information as much as possible.

- **Setup:** It takes as input a security parameter λ and outputs system public parameters $Params$.
- **KeyGen:** It takes as input a security parameter λ, system public parameters $Params$ and returns a signing-verification key pair (sk, pk).
- **Subliminal-Free Sign:** An interactive protocol between the sender and the warden. Given a message M, a signature σ is returned.
- **Verify:** It takes as input system public parameters $Params$, a public key pk and a signature message (M, σ). It returns 1 if and only if σ is a valid signature on message M.

3.2 Security Model

In the proposed scheme, the warden participates in the generation of a signature, hence the ability of the warden to forge a signature is enhanced. We regard the warden as the adversary. The formal definition of existential unforgeability against adaptively chosen messages attacks (EUF-CMA) is based on the following EUF-CMA game involving a simulator $\mathcal{S}$ and a forger $\mathcal{F}$:

1. **Setup:** $\mathcal{S}$ takes as input a security parameter λ, and runs the Setup algorithm. It sends the public parameters to $\mathcal{F}$.
2. **Query:** In addition to hash queries, $\mathcal{F}$ issues a polynomially bounded number of queries to the following oracles:
   - *Key generation oracle* $\mathcal{O}_{KeyGen}$: Upon receiving a key generation request, $\mathcal{S}$ returns a signing key.
   - *Signing oracle* $\mathcal{O}_{Sign}$: $\mathcal{F}$ submits a message M, and $\mathcal{S}$ gives $\mathcal{F}$ a signature σ.
3. **Forgery:** Finally, $\mathcal{F}$ attempts to output a valid forgery (M, σ) on some new message M, *i.e.*, a message on which F has not requested a signature. $\mathcal{F}$ wins the game if σ is valid.

The advantage of $\mathcal{F}$ in the EUF-CMA game, denoted by $\mathsf{Adv}(\mathcal{F})$, is defined as the probability that it wins.

Definition 1. (Existential Unforgeability) *A probabilistic algorithm $\mathcal{F}$ is said to (t, q_H, q_S, ϵ)-break a subliminal-free signature scheme if $\mathcal{F}$ achieves the advantage $\mathsf{Adv}(\mathcal{F}) \geq \epsilon$, when running in at most t steps, making at most q_H adaptive queries to the hash function oracle H, and requesting signatures on at most q_S adaptively chosen messages. A subliminal-free signature scheme is said to be (t, q_H, q_S, ϵ)-secure if no forger can (t, q_H, q_S, ϵ)-break it.*

4 Subliminal-Free Variant of Schnorr Signature

4.1 Construction

- **Setup:** Let (p, q, g) be a discrete logarithm triple associated with group $\mathbb{G}_{g,p}$. Let A be the sender of message $M \subseteq \{0,1\}^*$, B be the receiver of M and W be the warden. It chooses $t \in_R (1, q)$, returns t to W and computes $T = g^t \mod p$. Also, let H_0, H be two hash functions, where $H_0 : \{0,1\}^* \to \mathbb{G}_{g,p}$ and $H : \{0,1\}^* \times \mathbb{G}_{g,p} \to (1, q)$. Then, the public parameters are $Params = (p, q, g, H_0, H, T)$.
- **KeyGen:** It returns $x \in_R (1, q)$ as a secret key and the corresponding public key is $y = T^x \mod p$.
- **Subliminal-Free Sign**:
  1. W chooses two secret large integers c and d satisfying $cd = 1 \mod q$. Also, W chooses $k_w \in_R (1, q)$, thus $\gcd(k_w, q) = 1$. Then W computes $\alpha = g^{k_w c} \mod p$ and sends α to A.
  2. A chooses $k_a \in_R (1, q)$, thus $\gcd(k_a, q) = 1$. Then A computes $h_0 = H_0(M)$, $\beta = \alpha^{k_a h_0} \mod p$ and sends (h_0, β) to W.
  3. W computes $r = \beta^d = \alpha^{k_a h_0 d} = g^{k_a k_w h_0 cd} = g^{k_a k_w h_0} \mod p$, $v_1 = y^{k_w^{-1}} \mod p$, and sends (r, v_1) to A.
  4. A computes $e = H(M \parallel r)$, $f = e^x \mod p$ and $v_2 = g^{k_a h_0} \mod p$. Then A prepares a non-interactive zero knowledge proof that $DL_e(f) = DL_T(y)$ and sends (e, f, v_2) to W.
  5. W computes $u = k_w v_1^{-1} f^{-1} v_2^{-1} \mod p$, $\theta = u^{-1} t \mod p$ and sends θ to A.
  6. A computes s' and then sends (M, s') to W:

$$\begin{aligned} s' &= k_a h_0 + \theta \cdot (v_1^{-1} f^{-1} v_2^{-1}) \cdot xe \\ &= k_a h_0 + (u^{-1} t) \cdot (v_1^{-1} f^{-1} v_2^{-1}) \cdot xe \\ &= k_a h_0 + (v_2 f v_1 k_w^{-1} t) \cdot (v_1^{-1} f^{-1} v_2^{-1}) \cdot xe \\ &= k_a H_0(M) + k_w^{-1} xte \mod q. \end{aligned}$$

  7. *Sign:* Upon receiving (M, s'), W checks if $h_0 = H_0(M)$ and $e = H(M \parallel r)$. If not, W terminates the protocol, else W computes

$$s = k_w s' = k_a k_w H_0(M) + k_w k_w^{-1} xte = k_a k_w H_0(M) + xte \mod q.$$

  Then W sends the signature message $(M, (e, s))$ to B.

– **Verify**: After receiving the signature message $(M, (e, s))$, B computes

$$r' = g^s y^{-e} \mod p$$

and $e' = H(M \parallel r')$. B returns 1 if and only if $e = e'$.

4.2 Consistency of Our Construction

On one hand, if the signature message $(M, (e, s))$ is valid, we have $s = k_a k_w H_0(M) + xte \mod q$. Thus,

$$\begin{aligned} r' &= g^s y^{-e} = g^{k_a k_w H_0(M) + xte \mod q} y^{-e} \\ &= g^{k_a k_w H_0(M)} T^{xe} y^{-e} \\ &= g^{k_a k_w H_0(M)} y^e y^{-e} \\ &= g^{k_a k_w H_0(M)} \\ &= r \mod p, \end{aligned}$$

and then $e' = H(M \parallel r') = H(M \parallel r) = e$.

On the other hand, if $e = e'$, the signature message $(M, (e, s))$ is valid. Otherwise, we have

$$s \neq k_a k_w H_0(M) + xte \mod q$$

and then $r' \neq r$. However,

$$e' = H(M \parallel r') = H(M \parallel r) = e.$$

Thus, a collision of the hash function H is obtained, which is infeasible for a secure hash function.

5 Analysis of the Proposed Subliminal-Free Signature Scheme

5.1 Existential Unforgeability

Theorem 1. *If $\mathbb{G}_{g,p}$ is a (t', ϵ')-CDH group, then the proposed scheme is $(t, q_{H_0}, q_H, q_S, \epsilon)$-secure against existential forgery on adaptively chosen messages in the random oracle model, where*

$$t \geq t' - (q_H + 3.2q_S) \cdot C_{Exp}, \tag{1}$$

$$\epsilon \leq \epsilon' + q_S \cdot (q_{H_0} + q_S) 2^{-l_M} + q_S(q_H + q_S) 2^{-l_r} + q_H 2^{-l_q}, \tag{2}$$

where C_{Exp} denotes the cost of a modular exponentiation in group $\mathbb{G}_{g,p}$.

Proof. (sketch) Let $\mathcal{F}$ be a forger that $(t, q_{H_0}, q_H, q_S, \epsilon)$-breaks our proposed scheme. We construct a "simulator" algorithm $\mathcal{S}$ which takes $((p, q, g), (g^a, g^b))$ as inputs and runs $\mathcal{F}$ as a subroutine to compute the function $DH_{g,p}(g^a, g^b) = g^{ab}$ in t' steps with probability ϵ', which satisfy the Equalities (1) and (2). $\mathcal{S}$

makes the signer's verification key $y = g^a \mod p$ public, where the signing key a is unknown to $\mathcal{S}$. Aiming to translate $\mathcal{F}$'s possible forgery $(M, (e, s))$ into an answer to the function $DH_{g,p}(g^a, g^b)$, $\mathcal{S}$ simulates a running of the proposed scheme and answers $\mathcal{F}$'s queries. S uses $\mathcal{F}$ as a subroutine. Due to space limitation, we don't present the details here. ■

5.2 Subliminal-Freeness

It can be seen from the proposed scheme that the receiver B can only obtain the signature message $(M, (e, s))$ and temporary value r in addition to the verification public key y, thus it is necessary for the sender A to use e, s or r as a carrier when transmitting subliminal information.

In the following, we demonstrate that none of e, s and r can be controlled by A. On one hand, although the parameters $(\alpha, v_1, \theta) = (g^{k_w c}, y^{k_w^{-1}}, u^{-1}t) \mod p$ can be obtained by A, the secret exponents c, d and the secret parameters t, u are unknowable to him. Thus, A cannot obtain any information about k_w and g^{k_w}. Particularly, A knows nothing of k_w and g^{k_w} in the whole process of signing, hence the value of $s = k_w s' \mod p$ cannot be controlled by A. On the other hand, although the signer A computes $e = H(M \parallel r)$, nothing of k_w and g^{k_w} is available to him. Thus, the value of $r = g^{k_a k_w H_0(M)} \mod p$ cannot be controlled by A, and hence the value of e cannot be controlled. Note that if the value r generated by the warden W is not used by A in the Step 4, W can detect this fact in the Step 7 and terminate the protocol. Furthermore, if A attempts to directly compute k_w from g^{k_w}, he has to solve the discrete logarithm problem in group $GF^*(p)$, which is infeasible according to the intractability assumption of DLP.

Hence, we realize the complete subliminal-freeness of the subliminal channels existing in the random session keys in Schnorr signature scheme.

6 Conclusions and Future Work

In this paper, a subliminal-free protocol for Schnorr signature scheme is proposed. The proposed protocol completely closes the subliminal channels existing in the random session keys in Schnorr signature scheme. More strictly, it is completely subliminal-free in computational sense, and its security relies on the CDH assumption in the random oracle model. In addition, it is indispensable for the sender and the warden to cooperate with each other to sign a given message, and the warden is *honest-but-curious* and cannot forge a signature independently.

It would be interesting to construct subliminal-free signature schemes provably secure in the standard model.

Acknowledgements. We are grateful to the anonymous referees for their invaluable suggestions. This work is supported by the National Natural Science Foundation of China (No.61272457), the Nature Science Basic Research Plan in Shaanxi Province of China (No.2011JQ8042), and the Fundamental Research Funds for the Central Universities (Nos.K50511010001 and K5051201011). In particular, this work is supported by the Graduate Student Innovation Fund

of Xidian University (Research on key security technologies of large-scale data sharing in cloud computing).

References

1. Simmons, G.J.: The prisoner' problem and the subliminal channel. In: Advances in Cryptology-Crypto 1983, pp. 51–67. Plenum Press (1984)
2. Gupta, P.: Cryptography based digital image watermarking algorithm to increase security of watermark data. International Journal of Scientific & Engineering Research 3(9), 1–4 (2012)
3. Danezis, G., Kohlweiss, M., Rial, A.: Differentially Private Billing with Rebates. In: Filler, T., Pevný, T., Craver, S., Ker, A. (eds.) IH 2011. LNCS, vol. 6958, pp. 148–162. Springer, Heidelberg (2011)
4. Claycomb, W.R., Huth, C.L., Flynn, L., McIntire, D.M., Lewellen, T.B.: Chronological examination of insider threat sabotage: preliminary observations. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications 3(4), 4–20 (2012)
5. Choi, B., Cho, K.: Detection of insider attacks to the web server. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications 3(4), 35–45 (2012)
6. Lee, K., Lee, K., Byun, J., Lee, S., Ahn, H., Yim, K.: Extraction of platform-unique information as an identifier. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications 3(4), 85–99 (2012)
7. Chen, C.-L., Liao, J.-J.: A fair online payment system for digital content via subliminal channel. Electronic Commerce Research and Applications 10(3), 279–287 (2011)
8. Zhou, X., Yang, X., Wei, P., Hu, Y.: An anonymous threshold subliminal channel scheme based on elliptic curves cryptosystem. In: Computer-Aided Industrial Design and Conceptual Design, CAIDCD 2006, pp. 1–5 (November 2006)
9. Kim, K., Zhang, F., Lee, B.: Exploring signature schemes with subliminal channel. In: Symposium on Cryptography and Information Security 2003, pp. 245–250 (2003)
10. Yang, T.H.C.-L., Li, C.-M.: Subliminal channels in the identity-based threshold ring signature. International Journal of Computer Mathematics 86(5), 753–770 (2009)
11. Lin, D.-R., Wang, C.-I., Zhang, Z.-K., Guan, D.J.: A digital signature with multiple subliminal channels and its applications. Computers & Mathematics with Applications 60(2), 276–284 (2010); Advances in Cryptography, Security and Applications for Future Computer Science.
12. Troncoso, C., Danezis, G., Kosta, E., Balasch, J., Preneel, B.: Pripayd: Privacy-friendly pay-as-you-drive insurance. IEEE Transactions on Dependable and Secure Computing 8(5), 742–755 (2011)
13. Schnorr, C.P.: Efficient Identification and Signatures for Smart Cards. In: Brassard, G. (ed.) CRYPTO 1989. LNCS, vol. 435, pp. 239–252. Springer, Heidelberg (1990)
14. Desmedt, Y.: Simmons' protocol is not free of subliminal channels. In: Proceedings of 9th IEEE Computer Security Foundations Workshop, pp. 170–175 (1996)
15. Simmons, G.J.: Subliminal Communication Is Easy Using the DAS. In: Helleseth, T. (ed.) EUROCRYPT 1993. LNCS, vol. 765, pp. 218–232. Springer, Heidelberg (1994)

16. Xiangjun, X., Qingbo, L.: Construction of subliminal channel in id-based signatures. In: WASE International Conference on Information Engineering, ICIE 2009, vol. 2, pp. 159–162 (2009)
17. Xie, Y., Sun, X., Xiang, L., Luo, G.: A security threshold subliminal channel based on elliptic curve cryptosystem. In: Processing of IIHMSP 2008 International Conference on Intelligent Information Hiding and Multimedia Signal 2008, pp. 294–297 (2008)
18. Simmons, G.J.: The subliminal channels of the us digital signature algorithm (DSA). In: Advances in Cryptology-Cryptography, SPRC 1993, pp. 15–16 (1993)
19. Simmons, G.J.: An introduction to the mathematics of trust in security protocols. In: Proceedings of Computer Security Foundations Workshop VI, 1993, pp. 121–127 (June 1993)
20. Simmons, G.J.: Results concerning the bandwidth of subliminal channels. IEEE Journal on Selected Areas in Communications 16(4), 463–473 (1998)
21. Qingjun, C., Yuli, Z.: Subliminal channels in the NTRU and the subliminal-free methods. Wuhan University Journal of Natural Sciences 11, 1541–1544 (2006)
22. Sun, Y., Xu, C., Yu, Y., Yang, B.: Improvement of a proxy multi-signature scheme without random oracles. Computer Communications 34(3), 257–263 (2011); Special Issue of Computer Communications on Information and Future Communication Security.
23. Jadhav, M.V.: Effective detection mechanism for TCP based hybrid covert channels in secure communication. In: 2011 International Conference on Emerging Trends in Electrical and Computer Technology, ICETECT, pp. 1123–1128 (2011)
24. Yanai, N., Tso, R., Mambo, M., Okamoto, E.: A certificateless ordered sequential aggregate signature scheme secure against super adverssaries. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications 3(1), 30–54 (2012)
25. Simmons, G.J.: Subliminal channels: past and present. European Transactions on Telecommunications 5(4), 459–474 (1994)

A Block Cipher Mode of Operation with Two Keys

Yi-Li Huang, Fang-Yie Leu, Jung-Chun Liu, and Jing-Hao Yang

Department of Computer Science, TungHai University, Taiwan
{yifung,leufy,jcliu,g01350036}@thu.edu.tw

Abstract. In this paper, we propose a novel block cipher mode of operation (BCMO for short), named Output Protection Chain (OPC for short), which as a symmetric encryption structure is different from other existing BCMOs in that it employs two keys, rather than one key, to protect the output of the mode. The security threats of chosen-plaintext attacks on three existing common BCMOs, including the Cipher Feedback mode (CFB), the Output Feedback mode (OFB), and the Counter mode (CTR), are also analyzed. After that, we explain why the OPC mode (or simply the OPC) can effectively avoid chosen-plaintext attacks, and why its security level is higher than those of CFB, OFB, and CTR.

Keywords: Block cipher, Cipher Feedback mode, Output Feedback mode, Counter mode, Output Protection Chain mode, chosen-plaintext attack.

1 Introduction

When standard block cipher algorithms, like Data Encryption Standard (DES), Triple Data Encryption Algorithm (3DES), and Advanced Encryption Standard (AES), are used to encrypt a plaintext block, the size of the block should be the same as the length of the encrypting key (or called the ciphering block) L. If the size exceeds L, we have to divide the plaintext block into sub-blocks. Each is L in length. Several BCMOs defined by National Institute of Standards and Technology (NIST) have been widely adopted by different block cipher techniques [2]. Through the use of these BCMOs, these techniques can be then applied to many applications.

Generally, the standard BCMOs include the cipher Feedback mode (CFB for short), the Output Feedback mode (OFB for short) and the Counter mode (CTR for short), the characteristics of which are that they use only one key to encrypt multiple plaintext blocks, and the efficiencies of their block cipher algorithms are high [1]. Currently, different types of attacks on these BCMOs have been developed [3] [4], meaning the BCMOs have their own security problems. Therefore, in this study, we propose a novel BCMO, named Output Protection Chain (OPC for short), to solve the existing BCMOs' security problems. Two different structures of the OPC, named OPC-1 and OPC-2, have been developed to enhance the security levels of BCMOs. We will describe the two structures later.

2 Block Cipher Modes of Operation

Before describing operations of the CFB, OFB, and CTR, we first define the parameters used by them.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 392–398, 2013.
IFIP International Federation for Information Processing 2013

P_i : The i^{th} plaintext block to be encrypted, $1 \le i \le n$.
C_i : The i^{th} ciphertext block, $1 \le i \le n$.
Block Cipher Encryption (BCE) unit: According to [2], the standard BCE units are AES-128, AES-192, and AES-256. The function of a BCE unit is denoted by $E(I_P, K)$, in which the key K and the input I_p are used to encrypt a given plaintext block.
K: The block cipher key [2].
O_i: The output block produced by invoking the $E(I_p, K)$, $1 \le i \le n$.
cr: The counter, which is an input of the BCE unit of the CTR.
IV: Initialization Vector (IV for short), a random value employed by the CFB and OFB since they need an additional initial input.

The general rule in CFB is that $E(C_{i-1}, K)$ receives C_{i-1} and K as its inputs to generate O_i which is then XORed with P_i to produce C_i, $1 \le i \le n$, where $C_0 = IV$. The process can be formulated as follows.

$$C_i = P_i \oplus E(C_{i-1}, K) = P_i \oplus O_i \quad (1)$$

The encryption operations of the OFB are similar to those of the CFB. The difference is the inputs of $E(I_P, K)$. In the OFB, O_{i-1}, rather than C_{i-1}, is fed back to the BCE unit to generate O_i, $1 \le i \le n$, where $O_0 = IV$. It can be formulated as follows.

$$C_i = P_i \oplus E(O_{i-1}, K) = P_i \oplus O_i \quad (2)$$

The CTR encryption replaces the feedback operation employed by the CFB and OFB with a counter cr as one of the inputs of the BCE unit to generate O_i. The value of the counter used to generate O_i is $cr + i - 1$ where cr is the value adopted to produce O_i, $1 \le i \le n$. The formulas utilized to encrypt plaintext blocks of the CTR are as follows.

$$C_i = P_i \oplus E(cr + i - 1, K) \quad (3)$$

3 The Output Protection Chain (The OPC)

In this section, we describe how to encrypt plaintext blocks and decrypt ciphertext blocks in the proposed OPC structures, i.e., OPC-1 and OPC-2. We first define those parameters and functions invoked by the OPCs.

The definitions of P_i, C_i, BCE units, $E(I_p, K)$ and O_i, $1 \le i \le n$, are the same as those defined above. New parameters, operations and functions are defined below.

Key1: The block cipher key, the role of which is the same as K defined above.
$D(I_p, K)$: Function of the block decipher, in which the key K and an input I_P are used to decrypt a plaintext block from its ciphertext block.
G_i: The data block produced by $O_i \oplus P_i$, $1 \le i \le n$.
Key2: A key with the length the same as that of O_i. It is used to encrypt G_1 in the OPC-1, and O_1 in the OPC-2.
$+_2$: a binary adder, which is a logical operator defined in [5].
$-_2$: The Inverse operation of $+_2$.

3.1 The OPC-1

As shown in Fig.1, the general rule of the OPC-1 is that Key1 and G_{i-1} are input to the BCE unit to generate O_i, which is XORed with P_i to produce G_i. G_i is then binary-added with O_{i-1} to generate C_i, $1 \le i \le n$. The formulas derived are as follows.

$$C_i = [E(G_{i-1}, Key1) \oplus P_i] +_2 O_{i-1} = G_i +_2 O_{i-1} \tag{4}$$

where $G_0 = IV$ and $O_0 = Key2$. The decryption process as shown in Fig. 2 can be formulated as follows.

$$P_i = O_i \oplus G_i = E(G_{i-1}, Key1) \oplus (C_i -_2 O_{i-1}) \tag{5}$$

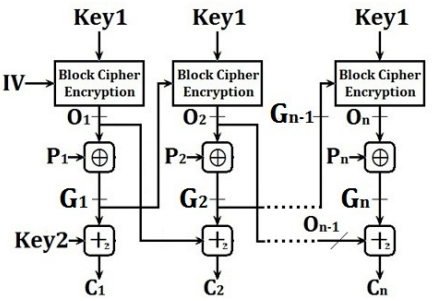


Fig. 1. The OPC-1 encryption

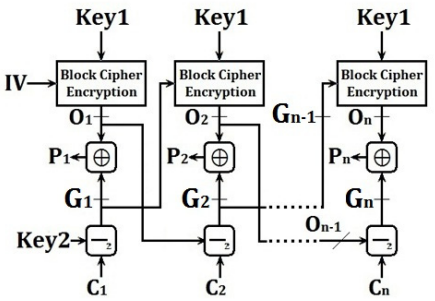


Fig. 2. The OPC-1 decryption

3.2 The OPC-2

The encryption process of the OPC-2 is shown in Fig. 3. The general rule is that P_i and Key1 are input to the BCE unit to generate O_i, which is XORed with O_{i-1} to generate C_i,$1 \le i \le n$. It can be formulated as follows.

$$C_i = O_i \oplus O_{i-1} = E(P_i, Key1) \oplus O_{i-1} \tag{6}$$

where $O_0 = Key2$. The decryption structure of the OPC-2 as shown in Fig. 4 is as follows. To decrypt C_i, one needs O_{i-1} to calculate O_i because $O_i = C_i \oplus O_{i-1}$, $1 \le i \le n$. P_i can be obtained by invoking the following formulas.

$$P_i = D(C_i \oplus O_{i-1}, Key1) = D(O_i, Key1) \tag{7}$$

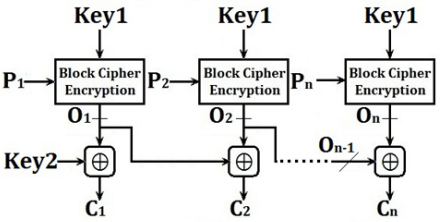


Fig. 3. The OPC-2 encryption

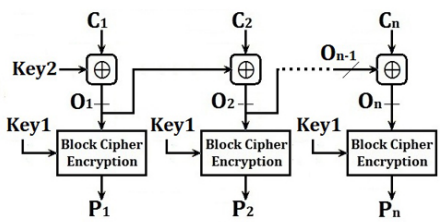


Fig. 4. The OPC-2 decryption

4 Security Analysis

The advantage of using BCMOs is that these BCMOs can enhance security of a single block's encryption. Even if the block cipher (e.g. DES) has been cracked, in order to improve the security level of a security system, one can apply the DES as the BCE unit to the BCMOs. We will analyze the security of BCMOs mentioned above in the following subsections.

4.1 Security of the CFB

To launch a chosen-plaintext attack, an attacker first inputs n different plaintext blocks, denoted by $P = \{P_1, P_2, \ldots, P_n\}$, to acquire a set of n ciphertext blocks, denoted by $C = \{C_1, C_2, \ldots, C_n\}$, where P_i is the i^{th} block of P, and C_i is the i^{th} block of C, $1 \le i \le n$. In the CFB, O_i can be derived from P_i and C_i since $O_i = P_i \oplus C_i$. If n is huge, the attacker can then collect sufficient $\langle C_{i-1}, O_i \rangle$ pairs, as the input and output of the BCE unit when encrypting P_i, to analyze the value of the key K.

4.2 Security of the OFB

For the OFB, we analyze its security based on two cases, one is that the IV can be chosen by users, and the other is cannot be chosen.

4.2.1 Attack on IV Able to be Chosen

In the OFB, O_i, $1 \le i \le n$, is only determined by IV and K. If IV can be chosen by users, the attacker can select the IV the same as the one chosen by a user, i.e., the victim, to encrypt their chosen-plaintext and calculate O_i by using $P_i \oplus C_i$.

Since K and the encryption algorithm of the BCE unit when encrypting different plaintext blocks are themselves the same, that means once the chosen IVs for encrypting two plaintexts are the same. When O_is of the BCE unit are acquired, the attacker can use an illegally intercepted C_i to search the corresponding O_i from all its collected $\langle C_i, O_i \rangle$ pairs to derive P_i without requiring breaking the key K of the BCE unit, since $P_i = C_i \oplus O_i$.

4.2.2 Attack on IV Unable to be Chosen

If the IV cannot be chosen, the security level of the OFB is higher. But it still faces the same security problem of the CFB. Like that in attacking the CFB, the attacker can first input a long plaintext, $P = \{P_1, P_2, \ldots, P_n\}$, to acquire the corresponding ciphertext, $C = \{C_1, C_2, \ldots, C_n\}$, so as to generate a set of $O = \{O_1, O_2, \ldots, O_n\}$ since $O_i = P_i \oplus C_i$.

If n is huge enough, the attacker can then collect sufficient $\langle O_{i-1}, O_i \rangle$ pairs to analyze the key K of its BCE unit. After that, when the attacker eavesdrops the messages delivered between the sender and receiver, and retrieves the IV, he/she can generate O to decrypt the intercepted C so as to obtain P since $P = C \oplus O$.

4.3 Security of the CTR

Like that in the OFB, the CTR can also be divided into two cases, i.e., a user can and cannot choose the value of cr.

4.3.1 Attack on CR Able to be Chosen

In the CTR encryption, O_is are determined only by cr and K in which cr is an incremental integer (will be transformed into a bit string) and K is a fixed key. The general rule is that given a chosen cr, $E(cr + i - 1, K)$ receives $cr + i - 1$ and K, $1 \le i \le n$, as its inputs to generate a set of output $O = \{O_1, O_2, \ldots, O_n\}$, $n \gg 1$, without requiring inputting any plaintext. On the other hand, for the chosen cr, the attacker can choose a plaintext $P = \{P_1, P_2, \ldots, P_n\}$ for the CTR to generate a ciphertext $C = \{C_1, C_2, \ldots, C_n\}$. After that, for each i, $1 \le i \le n$, $O_i = C_i \oplus P_i$. Then the attacker can acquire O.

If users can choose cr, then the attacker can also choose a cr the same as that of a user, i.e., the victim, to obtain O corresponding to this cr by using the above process. Now the attacker can decrypt the plaintext block P_i from the ciphertext block C_i intercepted from the user by using $P_i = C_i \oplus O_i$.

4.3.2 Attack on CR Unable to be Chosen

If the cr cannot be chosen, the attacker is still able to know the cr. Because cr is delivered together with C_is to the receiver with cr unencrypted [6], the attacker can obtain n input blocks, i.e., $I_p = \{cr, cr + 1, \ldots, cr + i - 1\}$, $n \gg 1$, from those messages carrying crs and C_is.

On the other hand, the attacker can input a set of plaintext blocks $P = \{P_1, P_2, \ldots, P_n\}$ to the CTR to obtain the corresponding ciphertext blocks, denoted by $C = \{C_1, C_2, \ldots, C_n\}$. He/she can then acquire a set of outputs of the BCE unit, denoted by $O = \{O_1, O_2, \ldots, O_n\}$, since $O_i = P_i \oplus C_i$, $1 \le i \le n$. As a result, the attacker can analyze the key K used by the BCE unit after collecting a large number of $\langle cr + i - 1, O_i \rangle$ pairs.

4.4 Security of the OPC-1

In Fig. 1, we use Key2 to protect G_1 and produce C_1, where$G_1 = P_1 \oplus O_1$. After that, O_i, $i > 1$, as the new Key2 of the next encryption round, is used to encrypt G_{i+1} to generate C_{i+1}. The advantage is that, when a large number of chosen-plaintext is input, C_i collected by the attacker is the one encrypted by O_{i-1} or Key2 (when $i = 1$). So there is no way for the attacker to decrypt G_is by using an inverse operation on C_is without knowing Key2 beforehand.

Moreover, G_i is fed back to generate O_{i+1}. The purpose is to increase the complexity of solving Key1. Also, G_i is encrypted by O_{i-1} or Key2, resulting in the fact that it is hard for the attacker to analyze the relationship between P_i and C_i, $1 \le i \le n$.

The shortcoming of the OPC-1 is that the plaintext is not encrypted by the BCE unit so the possibility for the attacker to decrypt P_i from C_i is still high since the attacker does not need to decrypt the BCE unit.

4.5 Security of the OPC-2

Fig. 3 shows the OPC-2 encryption, in which Key2 is used to encrypt O_1 so as to produce C_1. Meanwhile, O_1 is also used to encrypt O_2, i.e., $C_i = O_i \oplus O_{i-1}$, $1 < i \leq n$. As with the OPC-1, it is hard for the attacker to acquire O_i by decrypting C_i since O_i is encrypted by O_{i-1} or Key2, even he/she has collected a large number of ciphertext by inputting many chosen-plaintexts to the OPC-2.

If the attacker wishes to analyze the BCE unit of the OPC-2, he/she can input a long plaintext $P = \{P_1, P_2, \ldots, P_n\}$ of n plaintext blocks to the OPC-2 to generate the corresponding ciphertext $C = \{C_1, C_2, \ldots, C_n\}$. But before generating the set of output $O = \{O_1, O_2, \ldots, O_n\}$ of the BCE unit, he/she still needs to know Key2 because $O_1 = C_1 \oplus \text{Key2}$, and $O_i = C_i \oplus O_{i-1}$, $1 < i \leq n$. In our design, all encryption and decryption steps are dependent, so it is impossible to acquire O_1 without knowing Key2. Moreover, P_i is also protected by the BCE unit. With this, OPC-2 effectively strengthens the security level of O_i. As a result, it is hard for the attacker to collect sufficient information to analyze Key1 of the BCE unit.

5 Conclusions and Future Studies

In this paper, we describe the security drawbacks of the standard BCMOs, and propose the OPCs to improve the security level of a block ciphering system by protecting the outputs of its BCE unit, i.e., O_is, without the need of preventing the attacker from collecting P_is, C_is and their relationship. The purpose is avoiding the security system from being attacked by known or chosen-plaintext/ciphertext attacks.

However, in the OPC-2, the BCE unit must be invertible, e.g., DES, 3-DES, or AES. Otherwise, the plaintext P_i cannot be reverted from O_i. Since the encryption speeds of non-invertible algorithms are often short, and their encryption keys are difficult to crack, if one replaces the BCE unit of the CFB, OFB, CTR or OPC-1 with a non-invertible algorithm, the security levels and the processing performance of these BCMOs will be then higher than before. Therefore, in the future, we will apply non-invertible algorithms to the OPC-1 so as to propose a new BCMO with the security at least the same as or higher than those of the two OPCs.

Acknowledgements. This research was partially supported by TungHai University on GREENs project, and National Science Council, Taiwan, grants NSC 100-2221-E-029-018 and 101-2221-E-029-003-MY3.

References

1. Stallings, W.: Cryptography and Network Security: Principles and Practice, 5th edn. Prentice Hall (January 2010)
2. National Institute of Standards and Technology, NIST Special Publication 800-38A, Recommendation for Block Cipher Modes of Operation Methods and Techniques (December 2001)

3. Hudde, H.: Building Stream Ciphers from Block Ciphers and their Security. Seminararbeit Ruhr-Universität Bochum (February 2009), http://imperia.rz.rub.de:9085/imperia/md/content/seminare/itsws08_09/hudde.pdf
4. Wang, D., Lin, D., Wu, W.: Related-Mode Attacks on CTR Encryption Mode. International Journal of Network Security 4(3), 282–287 (2007)
5. Huang, Y.F., Leu, F.Y., Chiu, C.H., Lin, I.L.: Improving Security Levels of IEEE802.16e Authentication by Involving Diffie-Hellman PKDS. Journal of Universal Computer Science 17(6), 891–911 (2011)
6. Lipmaa, H., Rogaway, P., Wagner, D.: Comments to NIST concerning AES Modes of Operations: CTR-Mode Encryption (2000), http://csrc.nist.gov/

On the Security of an Authenticated Group Key Transfer Protocol Based on Secret Sharing

Ruxandra F. Olimid

Department of Computer Science, University of Bucharest, Romania
ruxandra.olimid@fmi.unibuc.ro

Abstract. Group key transfer protocols allow multiple parties to share a common secret key. They rely on a mutually trusted key generation center (KGC) that selects the key and securely distributes it to the authorized participants. Recently, Sun et al. proposed an authenticated group key transfer protocol based on secret sharing that they claim to be secure. We show that this is false: the protocol is susceptible to insider attacks and violates known key security. Finally, we propose a countermeasure that maintains the benefits of the original protocol.

Keywords: group key transfer, secret sharing, attack, cryptanalysis.

1 Introduction

Confidentiality represents one of the main goals of secure communication. It assures that the data are only accessible to authorized parties and it is achieved by encryption. In case of symmetric cryptography, the plaintext is encrypted using a secret key that the sender shares with the qualified receiver(s). Under the assumption that the system is secure, an entity that does not own the private key is unable to decrypt and thus the data remain hidden to unauthorized parties.

The necessity of a (session) key establishment phase before the encrypted communication starts is immediate: it allows the authorized parties to share a common secret key that will be used for encryption.

Key establishment protocols divide into key transfer protocols - a mutually trusted key generation center (KGC) selects a key and securely distributes it to the authorized parties - and key agreement protocols - all qualified parties are involved in the establishment of the secret key. The first key transfer protocol was published by Needham and Schroeder in 1978 [11], two years after Diffie and Hellman had invented the public key cryptography and the notion of key agreement [4].

The previous mentioned protocols restrict to the case of two users. As a natural evolution, *group (or conference) key establishment protocols* appeared a few years latter. Ingemarsson et al. introduced the first key transfer protocol that permits to establish a private key between multiple parties [8]. Their protocol generalizes the Diffie-Hellman key exchange.

A general construction for a key transfer protocol when KGC shares a long-term secret with each participant is straightforward: KGC generates a fresh

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 399–408, 2013.

key and sends its encryption (under the long-term secret) to each authorized party. The qualified users decrypt their corresponding ciphertext and find the session secret key, while the unqualified users cannot decrypt and disclose the session key. KGC performs n encryptions and sends n messages, where n is the number of participants. Therefore, the method becomes inefficient for large groups.

Secret sharing schemes are used in group key transfer protocols to avoid such disadvantages. A secret sharing scheme splits a secret into multiple shares so that only authorized sets of shares may reconstruct the secret. Blakley [1] and Shamir [14] independently introduce secret sharing schemes as key management systems. The particular case when all shares are required for reconstruction is called *all-or-nothing secret sharing scheme*; the particular case when at least k out of n shares are required for reconstruction is called *(k,n)-threshold secret sharing scheme*.

Various group key establishment protocols based on secret sharing schemes exist in the literature. Blom proposed an efficient key transfer protocol in which every two users share a common private key that remains hidden when less than k users cooperate [2]. Blundo et al. generalized Blom's protocol by allowing any t users to share a private key, while it remains secure for a coalition of up to k users [3]. Fiat and Naor improved the construction even more by permitting any subset of users to share a common key in the same conditions [5]. Pieprzyk and Li gave a couple of group key agreement protocols based on Shamir's secret scheme [9,12]. Recently, Harn and Lin also used Shamir's scheme to construct a key transfer protocol [6], which was proved to be insecure and further adjusted [10]. Some other examples from literature include Sáez's protocol [13] (based on a family of vector space secret sharing schemes), Hsu et al.'s protocol [7] (based on linear secret sharing schemes) and Sun et al's group key transfer protocol [15], which we will refer for the rest of this paper.

Sun et al. claim that their construction is secure and provides several advantages: each participant stores a single long-term secret for multiple sessions, computes the session key by a simple operation and the protocol works within dynamic groups (i.e. members may leave or join the group). We demonstrate that they are wrong. First, we show that the protocol is susceptible to insider attacks: any qualified group member may recover a session key that he is unauthorized to know. Second, we prove that the protocol violates known key security: any attacker who gains access to one session key may recover any other key. We propose an improved version of Sun et al's group key transfer protocol that stands against both attacks and achieves the benefits claimed in the original work.

The paper is organized as follows. The next section contains the preliminaries. Section 3 describes Sun et al.'s authenticated group key transfer protocol. Section 4 introduces the proposed attacks. In Section 5 we analyze possible countermeasures. Section 6 concludes.

2 Preliminaries

2.1 Security Goals

Group key transfer protocols permit multiple users to share a common private key by using pre-established secure communication channels with a trusted KGC, which is responsible to generate and distribute the key. Each user registers to KGC for subscribing to the key distribution service and receives a long-term secret, which he will later use to recover the session keys.

We will briefly describe next the main security goals that a group key transfer protocol must achieve: *key freshness*, *key confidentiality*, *key authentication*, *entity authentication*, *known key security* and *forward secrecy*.

Key freshness ensures the parties that KGC generates a random key that has not been used before. Unlike key agreement protocols, the users are not involved in the key generation phase, so the trust assumption is mandatory.

Key confidentiality means that a session key is available to authorized parties only. Adversaries are categorized into two types: insiders - that are qualified to recover the session key - and outsiders - that are unqualified to determine the session key. A protocol is susceptible to insider attacks if an insider is able to compute secret keys for sessions he is unauthorized for. Similarly, it is vulnerable to outsider attacks if an outsider is capable to reveal any session key.

Key authentication assures the group members that the key is distributed by the trusted KGC and not by an attacker. It may also stand against a replay attack: no adversary can use a previous message originated from KGC to impose an already compromised session key.

Entity authentication confirms the identity of the users involved in the protocol, so that an attacker cannot impersonate a qualified principal to the KGC.

Known key security imposes that a compromised session key has no impact on the confidentiality of other session keys: even if an adversary somehow manages to obtain a session key, all the other past and future session keys remain hidden.

Forward secrecy guarantees that even if a long-term secret is compromised, this has no impact on the secrecy of the previous session keys.

2.2 Secret Sharing

A secret sharing scheme is a method to split a secret into multiple shares, which are then securely distributed to the participants. The secret can be recovered only when the members of an authorized subset of participants combine their shares together. The set of all authorized subsets is called the *access structure*. The access structure of a (k, n) *threshold secret sharing scheme* consists of all sets whose cardinality is at least k. The access structure of an *all-or-nothing secret sharing scheme* contains only one element: the set of all participants.

Generally, a secret sharing scheme has 3 phases: *sharing* (a dealer splits the secret into multiple parts, called *shares*), *distribution* (the dealer securely transmits the shares to the parties) and *reconstruction* (an authorized group of parties put their shares together to recover the secret).

Group key establishment protocols use secret sharing schemes due to the benefits they introduce: decrease computational and transmission costs, represent a convenient way to differentiate between principals and their power within the group, permits delegation of shares, accepts cheating detection, permits the sizing of the group, etc. [12]. For more information, the reader may refer to [12].

2.3 Discrete Logarithm Assumption

Let G be a cyclic multiplicative group of order p with $g \in G$ a generator.

The Discrete Logarithm Assumption holds in G if given g^a, any probabilistic polynomial-time adversary $\mathcal{A}$ has a negligible probability in computing a:

$$Adv_{\mathcal{A}} = Pr[\mathcal{A}(p, g, g^a) = a] \leq negl(k) \tag{1}$$

where $a \in \mathbb{Z}_p^*$ is random and k is the security parameter.

3 Sun et al.'s Group Key Transfer Protocol

Let n be the size of the group of participants, $\mathcal{U} = \{U_1, U_2, ..., U_n\}$ the set of users, G a multiplicative cyclic group of prime order p with $g \in G$ as a generator, H a secure hash function.

Sun et al. base their protocol on a *derivative secret sharing*, which we describe next and briefly analyze in Section 5:

Derivative Secret Sharing [15]
Phase 1: Secret Sharing
The dealer splits the secret $S \in G$ into two parts n times:

$$S = s_1 + s_1' = s_2 + s_2' = ... = s_n + s_n' \tag{2}$$

Phase 2: Distribution
The dealer sends the share $s_i' \in G$ to $U_i \in \mathcal{U}$ via a secure channel.
Phase 3: Reconstruction
1. The dealer broadcasts the shares $s_1, s_2, ..., s_n$ at once, when the users want to recover the secret S.
2. Any user $U_i \in \mathcal{U}$ reconstructs the secret as:

$$S = s_i' + s_i \tag{3}$$

Next, we review Sun et. al's group key transfer protocol, which we prove vulnerable in Section 4:

Sun et al.'s Group Key Transfer Protocol [15]
Phase 1: User Registration
During registration, KGC shares a long-term secret $s_i' \in G$ with each user $U_i \in \mathcal{U}$.

Phase 2: Group Key Generation and Distribution
1. A user, called the initiator, sends a group key distribution request that contains the identity of the qualified participants for the current session $\{U_1, U_2, ..., U_t\}$ to KGC.[1]
2. KGC broadcasts the received list as a response.
3. Each member $U_i, i = 1, \ldots, t$ that identifies himself in the list sends a random challenge $r_i \in \mathbb{Z}_p^*$ to KGC.
4. KGC randomly selects $S \in G$ and invokes the derivative secret sharing scheme to split S into two parts t times such that $S = s_1 + s_1' = s_2 + s_2' = ... = s_t + s_t'$. He computes the session private key as $K = g^S$, t messages $M_i = (g^{s_i+r_i}, U_i, H(U_i, g^{s_i+r_i}, s_i', r_i)), i = 1, \ldots, t$ and $Auth = H(K, g^{s_1+r_1}, ..., g^{s_t+r_t}, U_1, ..., U_t, r_1, ..., r_t)$. At last, KGC broadcasts $(M_1, ..., M_t, Auth)$ as a single message.
5. After receiving M_i and $Auth$, each user U_i, $i = 1, \ldots, t$ computes $h = H(U_i, g^{s_i+r_i}, s_i', r_i)$ using $g^{s_i+r_i}$ from M_i, s_i' the long-term secret and r_i as chosen in step 3. If h differs from the corresponding value in M_i, the user aborts; otherwise, he computes $K' = g^{s_i'} \cdot g^{s_i+r_i}/g^{r_i}$ and checks if $Auth = H(K', g^{s_1+r_1}, ..., g^{s_t+r_t}, U_1, ..., U_t, r_1, ..., r_t)$. If not, he aborts; otherwise, he consider K' to be the session key originated from KGC and returns a value $h_i = H(s_i', K', U_1, ..., U_t, r_1, ...r_t)$ to KGC.
6. KGC computes $h_i' = H(s_i', K, U_1, ..., U_t, r_1, ...r_t)$ using his own knowledge on s_i' and checks if h_i' equals h_i, certifying that all users possess the same key.

The authors claim that their construction is secure under the discrete logarithm assumption and has multiple advantages: each participant needs to store only one secret share for multiple sessions (the long-term secret $s_i', i = 1, \ldots, n$), the dynamic of the group preserves the validity of the shares (if a user leaves or joins the group there is no need to update the long-term secrets), each authorized user recovers the session key by a simple computation. Unlike their claim, in the next section we prove that the protocol is insecure.

4 The Proposed Attacks

We show that Sun et al.'s protocol is insecure against insider attacks and violates known key security.

4.1 Insider Attack

Let $U_a \in \mathcal{U}$ be an authorized user for a session (k_1), s_a' his long-term secret, $\mathcal{U}_{(k_1)} \subseteq \mathcal{U}$ the qualified set of participants of the session, $(g^{s_{i_{(k_1)}}+r_{i_{(k_1)}}})_{U_i \in \mathcal{U}_{(k_1)}}$ the values that were broadcast as part of $(M_i)_{U_i \in \mathcal{U}_{(k_1)}}$ in step 4 and $K_{(k_1)} = g^{S_{(k_1)}}$ the session key.

[1] Without loss of generality, any subset of $\mathcal{U}$ can be expressed as $\{U_1, U_2, ..., U_t\}$ by reordering.

The participant U_a is qualified to determine (k_1) session key as:

$$K_{(k_1)} = \frac{g^{s'_a} \cdot g^{s_{a(k_1)}+r_{a(k_1)}}}{g^{r_{a(k_1)}}} \tag{4}$$

Since $g^{s_{i(k_1)}+r_{i(k_1)}}$ and $r_{i(k_1)}$ are public, he is able to compute $g^{s'_i}$, for all $U_i \in \mathcal{U}_{(k_1)}$:

$$g^{s'_i} = \frac{K_{(k_1)} \cdot g^{r_{i(k_1)}}}{g^{s_{i(k_1)}+r_{i(k_1)}}} \tag{5}$$

Suppose that U_a is unauthorized to recover (k_2) session key, $(k_2) \neq (k_1)$. However, he can eavesdrop the exchanged messages. Therefore, he is capable to determine

$$g^{s_{j(k_2)}} = \frac{g^{s_{j(k_2)}+r_{j(k_2)}}}{g^{r_{j(k_2)}}} \tag{6}$$

for all $U_j \in \mathcal{U}_{(k_2)}$, where $\mathcal{U}_{(k_2)} \subseteq \mathcal{U}$ is the qualified set of parties of the session (k_2).

We assume that there exists a participant $U_b \in \mathcal{U}_{(k_1)} \bigcap \mathcal{U}_{(k_2)}$ that is qualified for both sessions (k_1) and (k_2). The inside attacker U_a can find the key $K_{(k_2)}$ of the session (k_2) as:

$$K_{(k_2)} = g^{s'_b} \cdot g^{s_{b(k_2)}} = g^{s'_b+s_{b(k_2)}} \tag{7}$$

In conclusion, an insider can determine any session key under the assumption that at least one mutual authorized participant for both sessions exists, which is very likely to happen.

The attack also stands if there is no common qualified user for the two sessions, but there exists a third one (k_3) that has a mutual authorized party with each of the former sessions. The extension is straightforward: let $U_{1,3}$, $U_{2,3}$ be the common qualified parties for sessions (k_1) and (k_3), respectively (k_2) and (k_3). U_a computes the key $K_{(k_3)}$ as in the proposed attack due to the common authorized participant $U_{1,3}$. Once he obtains the key $K_{(k_3)}$, he mounts the attack again for sessions (k_3) and (k_2) based on the common party $U_{2,3}$ and gets $K_{(k_2)}$.

The attack extends in chain: the insider U_a reveals a session key $K_{(k_x)}$ if he is able to build a chain of sessions $(k_1)...(k_x)$, where (k_i) and (k_{i+1}) have at least one common qualified member $U_{i,i+1}$, $i = 1, \ldots, x-1$ and U_a is authorized to recover the key $K_{(k_1)}$.

4.2 Known Key Attack

Suppose an attacker (insider or outsider) owns a session key $K_{(k_1)}$. We also assume that he had previously eavesdropped values $r_{i_{(k_1)}}$ in step 3 and $g^{s_{i(k_1)}+r_{i(k_1)}}$ in step 4 of session (k_1) so that for all $U_i \in \mathcal{U}_{(k_1)}$ he determines:

$$g^{s_{i(k_1)}} = \frac{g^{s_{i(k_1)}+r_{i(k_1)}}}{g^{r_{i(k_1)}}} \tag{8}$$

Because the session key $K_{(k_1)}$ is exposed, he can also compute $g^{s_i'}$, for all $U_i \in \mathcal{U}_{(k_1)}$:

$$g^{s_i'} = \frac{K_{(k_1)}}{g^{s_{i(k_1)}}} \tag{9}$$

Let (k_2) be any previous or future session that has at least one common qualified participant U_b with (k_1), i.e. $U_b \in \mathcal{U}_{(k_1)} \bigcap \mathcal{U}_{(k_2)}$. As before, the attacker eavesdrops $r_{b_{(k_2)}}$ and $g^{s_{b_{(k_2)}}+r_{b_{(k_2)}}}$ and computes

$$g^{s_{b_{(k_2)}}} = \frac{g^{s_{b_{(k_2)}}+r_{b_{(k_2)}}}}{g^{r_{b_{(k_2)}}}} \tag{10}$$

The attacker can now recover the key $K_{(k_2)}$:

$$K_{(k_2)} = g^{s_b'} \cdot g^{s_{b_{(k_2)}}} = g^{s_b'+s_{b_{(k_2)}}} \tag{11}$$

The attack may also be mount in chain, similar to the insider attack. We omit the details in order to avoid repetition.

The first attack permits an insider to reveal any session key he was unauthorized for, while the second permits an attacker (insider or outsider) to disclose any session key under the assumption that a single one has been compromised.

In both cases, the attacker computes the session key as the product of two values: g^{s_b} (disclosed only by eavesdropping) and $g^{s_b'}$ (revealed by eavesdropping when a session key is known). We remark that the attacker is unable to determine the long-term secret s_b' if the discrete logarithm assumption holds, but we emphasize this does not imply that the protocol is secure, since the adversary's main objective is to find the session key that can be disclosed without knowing s_b'.

5 Countermeasures

Sun et al.'s group key agreement protocol fails because values $g^{s_i'}$, $i = 1, \ldots, n$ are maintained during multiple sessions. We highlight that the derivative secret sharing scheme suffers from a similar limitation caused by the usage of the long-term secrets s_i', $i = 1, \ldots, n$ during multiple sessions: any entity that discloses a secret S determines the values $s_i' = S - s_i$ by eavesdropping s_i, $i = 1, \ldots, n$ and uses them to reveal other shared secrets.

A trivial modification prevents the proposed attacks: KGC replaces the values s_i' at the beginning of each session. This way, even if the attacker determines $g^{s_i'}$ in one round, the value becomes unusable. However, this introduces a major drawback: KGC must share a secure channel with each user for any round. If this were the case, KGC could have just sent the secret session key via the secure channel and no other protocol would have been necessary. In conclusion, the group key transfer protocol would have become useless.

Another similar solution exists: during the user registration phase an ordered set of secrets is shared between KGC and each user. For each session, the corresponding secret in the set is used in the derivative secret sharing. Although this

solution has the benefit that it only requires the existence of secure channels at registration, it introduces other disadvantages: each user must store a considerable larger quantity of secret information, the protocol can run for at most a number of times equal to the set cardinality, KGC must broadcast the round number so that participants remain synchronized.

We propose next a countermeasure inspired by the work of Pieprzyk and Li [12]. The main idea consist of using a different public value $\alpha \in G$ to compute the session key $K = \alpha^S$ for each round.

The Improved Version of the Group Key Transfer Protocol

Phase 1: User Registration
During registration, KGC shares a long-term secret $s'_i \in G$ with each user $U_i \in \mathcal{U}$.
Phase 2: Group Key Generation and Distribution
1. A user, called the initiator, sends a group key distribution request that contains the identity of the qualified participants for the current session $\{U_1, U_2, ..., U_t\}$ to KGC.
2. KGC broadcasts the received list as a response.
3. Each member $U_i, i = 1, \ldots, t$ that identifies himself in the list sends a random challenge $r_i \in \mathbb{Z}_p^*$ to KGC.
4. KGC randomly selects $S \in G$ and invokes the derivative secret sharing scheme to split S into two parts t times such that $S = s_1 + s'_1 = s_2 + s'_2 = ... = s_t + s'_t$. He chooses $\alpha \in G$ at random, computes the session private key as $K = \alpha^S$, t messages $M_i = (\alpha^{s_i+r_i}, U_i, H(U_i, \alpha^{s_i+r_i}, s'_i, r_i, \alpha)), i = 1, \ldots, t$ and $Auth = H(K, \alpha^{s_1+r_1}, ..., \alpha^{s_t+r_t}, U_1, ..., U_t, r_1, ..., r_t, \alpha)$. At last, KGC broadcasts $(M_1, ..., M_t, Auth, \alpha)$ as a single message.
5. After receiving M_i, $Auth$ and α, each user U_i, $i = 1, \ldots, t$ computes $h = H(U_i, \alpha^{s_i+r_i}, s'_i, r_i, \alpha)$ using $\alpha^{s_i+r_i}$ from M_i, s'_i the long-term secret and r_i as chosen in step 3. If h differs from the corresponding value in M_i, the user aborts; otherwise, he computes $K' = \alpha^{s'_i} \cdot \alpha^{s_i+r_i} / \alpha^{r_i}$ and checks if $Auth = H(K', \alpha^{s_1+r_1}, ..., \alpha^{s_t+r_t}, U_1, ..., U_t, r_1, ..., r_t, \alpha)$. If not, he aborts; otherwise, he consider K' to be the session key originated from KGC and returns a value $h_i = H(s'_i, K', U_1, ..., U_t, r_1, ...r_t, \alpha)$ to KGC.
6. KGC computes $h'_i = H(s'_i, K, U_1, ..., U_t, r_1, ...r_t, \alpha)$ using his own knowledge on s'_i and checks if h'_i equals h_i, certifying that all users possess the same key.

The countermeasure eliminates both attacks. Under the discrete logarithm assumption, a value ${\alpha_{(k_1)}}^{s'_i}$ from a session (k_1) can no longer be used to compute a session key $K_{(k_2)} = \alpha_{(k_2)}^{s'_i + s_{i_{(k_2)}}}$ with $(k_2) \neq (k_1)$. The values α are authenticated to originate from KGC so that an attacker cannot impersonate the KGC and use a suitable value (for example $\alpha_{(k_2)} = \alpha_{(k_1)}^a$ for a known a).

We remark that the modified version of the protocol maintains all the benefits of the original construction and preserves the computational cost, while the transmission cost increases negligible. However, we admit that it conserves a

weakness of the original protocol: cannot achieve forward secrecy. Any attacker that obtains a long-term secret becomes able to compute previous keys of sessions he had eavesdropped before. The limitation is introduced by construction because the long-term secret is directly used to compute the session key.

6 Conclusions

Sun et al. recently introduced an authenticated group key transfer protocol based on secret sharing [15], which they claimed to be efficient and secure. We proved that they are wrong: the protocol is vulnerable to insider attacks and violates known key security. We improved their protocol by performing a slight modification that eliminates the proposed attacks and maintains the benefits of the original work.

Acknowledgments. This paper is supported by the Sectorial Operational Program Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number SOP HDR/107/1.5/S/82514.

References

1. Blakley, G.: Safeguarding cryptographic keys. In: Proceedings of the 1979 AFIPS National Computer Conference, pp. 313–317. AFIPS Press (1979)
2. Blom, R.: An Optimal Class of Symmetric Key Generation Systems. In: Beth, T., Cot, N., Ingemarsson, I. (eds.) EUROCRYPT 1984. LNCS, vol. 209, pp. 335–338. Springer, Heidelberg (1985)
3. Blundo, C., De Santis, A., Herzberg, A., Kutten, S., Vaccaro, U., Yung, M.: Perfectly-Secure Key Distribution for Dynamic Conferences. In: Brickell, E.F. (ed.) CRYPTO 1992. LNCS, vol. 740, pp. 471–486. Springer, Heidelberg (1993)
4. Diffie, W., Hellman, M.E.: New directions in cryptography. IEEE Transactions on Information Theory 22(6), 644–654 (1976)
5. Fiat, A., Naor, M.: Broadcast Encryption. In: Stinson, D.R. (ed.) CRYPTO 1993. LNCS, vol. 773, pp. 480–491. Springer, Heidelberg (1994)
6. Harn, L., Lin, C.: Authenticated group key transfer protocol based on secret sharing. IEEE Trans. Comput. 59(6), 842–846 (2010)
7. Hsu, C., Zeng, B., Cheng, Q., Cui, G.: A novel group key transfer protocol. Cryptology ePrint Archive, Report 2012/043 (2012)
8. Ingemarsson, I., Tang, D.T., Wong, C.K.: A conference key distribution system. IEEE Transactions on Information Theory 28(5), 714–719 (1982)
9. Li, C.H., Pieprzyk, J.: Conference Key Agreement from Secret Sharing. In: Pieprzyk, J.P., Safavi-Naini, R., Seberry, J. (eds.) ACISP 1999. LNCS, vol. 1587, pp. 64–76. Springer, Heidelberg (1999)
10. Nam, J., Kim, M., Paik, J., Jeon, W., Lee, B., Won, D.: Cryptanalysis of a Group Key Transfer Protocol Based on Secret Sharing. In: Kim, T.-h., Adeli, H., Slezak, D., Sandnes, F.E., Song, X., Chung, K.-i., Arnett, K.P. (eds.) FGIT 2011. LNCS, vol. 7105, pp. 309–315. Springer, Heidelberg (2011)

11. Needham, R.M., Schroeder, M.D.: Using encryption for authentication in large networks of computers. Commun. ACM 21(12), 993–999 (1978)
12. Pieprzyk, J., Li, C.H.: Multiparty key agreement protocols. In: IEEE Proceedings - Computers and Digital Techniques, pp. 229–236 (2000)
13. Sáez, G.: Generation of key predistribution schemes using secret sharing schemes. Discrete Applied Mathematics 128(1), 239–249 (2003)
14. Shamir, A.: How to share a secret. Commun. ACM 22(11), 612–613 (1979)
15. Sun, Y., Wen, Q., Sun, H., Li, W., Jin, Z., Zhang, H.: An authenticated group key transfer protocol based on secret sharing. Procedia Engineering 29, 403–408 (2012)

Modified Efficient and Secure Dynamic ID-Based User Authentication Scheme

Toan-Thinh Truong[1], Minh-Triet Tran[1], and Anh-Duc Duong[2]

[1] University of Science, VNU-HCM
{ttthinh,tmtriet}@fit.hcmus.edu.vn
[2] University of Information Technology, VNU-HCM
daduc@uit.edu.vn

Abstract. Communication is necessary operations in wireless environments. Therefore, we must have a secure remote authentication to defend transactions against illegitimate adversaries in such risky channel. Smart card is one of methods that many schemes used due to its convenience. Recently, Khurram Khan has proposed an enhancement scheme using smart card to recover some pitfalls in Wang et al.'s scheme. They claimed that their scheme remedy those security flaws. Nevertheless, we point out that Khan et al.'s scheme cannot protect user's anonymity. Besides, it does not achieve secret key forward secrecy and cannot resist denial of service attack due to values stored in server's database. Consequently, we present an improvement to their scheme to isolate such problems.

Keywords: Mutual authentication, Dynamic identity, Anonymity.

1 Introduction

In network environment, users can access services via different devices, for example, PC, laptop, mobile phone. Communication between those devices and services can be operated with many network technologies, such as, wireless, 3G.

There are many methods of constructing secure authentication. In 1981, Lamport [1] is the first person applying cryptographic hash function in authentication. Later, many author also use this approach in their protocols. Typically, there are protocols of Cheng-Chi Lee [2] and Jau-Ji Shen [3]. These authors follow Lamport's approach with slight difference that they use identity to authenticate instead of password table. In 2004, Das et al proposed a scheme [4]. Their scheme has three main advantages. Firstly, it allows users to change password freely. Moreover, it does not maintain a verification table which is used to check login message. Finally, the scheme's security is based on secure one-way hash function. In 2005, I-En Liao [5] discovered Das's scheme not only cannot resist some basic kinds of attacks such as password-guessing but also do not provide mutual authentication. Furthermore, in Das's scheme, password is transmitted in plain-text form at registration phase. Therefore, it is easy to be stolen by server. In I-En Liao's scheme, author use hash function with password before transmitting it to server. So, even server do not know actual user's password. Nonetheless,

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 409–415, 2013.

with hash function, I-En Liao's scheme is also vulnerable to password-guessing attack. Consequently, E-J Yoon proposed an improvement [6] to I-En Liao's. In [6], author utilizes random value with password when hashing. So, this causes attacker not to be able to guess true user's password. Recently, Khuram Khan with E-J Yoon's approach devised a protocol [7]. In [7], authors also distribute common secret information to all users and use timestamp to resist replay attack. They also claimed their scheme can protect user's anonymity. In this paper, we prove that their scheme has inability to defend anonymity. Furthermore, it also cannot achieve secret key forward secrecy and does not withstand denial of service attack due to values stored in database's server. Ultimately, we propose an improved version to recover problems mentioned.

The remainder of this paper is organized as follows: section 2 quickly reviews Khan's scheme and discusses its weaknesses. Then, our proposed scheme is presented in section 3, while section 4 discusses the security and efficiency of the proposed scheme. Our conclusions are presented in section 5.

2 Review and Cryptanalysis of Khurram Khan's Scheme

In this section, we review Khan's scheme [7] and show his scheme cannot obtain secret key forward secrecy and doesn't protect user's anonymity.

2.1 Review of Khurram Khan's Scheme

Their scheme includes five phases. Some important notations are listed as follow:

- U_i, S: Qualified user & remote server.
- pw_i, $h(.)$: Unique password of U_i and one-way hash function.
- x, y: The first & the second secret keys of the remote server.
- T, N: The timestamp & the number of times a user registers.
- SC, SK: The smart card & the session key.
- $\oplus$, $\|$: The exclusive-or operation & concatenation operation.

Registration Phase. U_i submits ID_i & $h(r \parallel pw_i)$ to S via a secure channel, where r is a random value chosen by user. Then, S performs the following steps.

1. S checks U_i's registration credentials & checks ID_i's existence. If it already exists & N is equal to 0, S requests U_i to choose another ID_i. Otherwise, S computes $J = h(x \parallel ID_U)$ where $ID_U = (ID_i \parallel N)$ and $L = J \oplus RPW$.
2. S issues SC with $\{L, y\}$ to U_i over secure channel. Then, U_i stores r in SC.

Login Phase. After receiving SC from S, U_i uses it to login into S.

1. U_i inserts SC into another terminal's card-reader. Then he keys pw_i & ID_i.
2. SC computes $RPW = h(r \parallel pw_i)$ & $J = L \oplus RPW$. Then, SC acquires T_i & computes $C_1 = h(T_i \parallel J)$.
3. SC generates a random value d & computes $AID_i = ID_i \oplus h(y \parallel T_i \parallel d)$. Finally, SC sends login message $m = \{AID_i, T_i, d, C_1\}$ to S.

Authentication Phase. S authenticates the users login request.

1. Verifies the validity of time interval between T_i & $T^{'}$. If ($T^{'}$ - T_i) $\geq \Delta T$, where ΔT denotes the expected valid time interval for transmission delay. If this holds, then S rejects & terminates the session.
2. Computes $ID_i = AID_i \oplus h(y \parallel T_i \parallel d)$ & checks if ID_i is valid, otherwise terminates the operation. Then S checks N in database & computes $ID_U = (ID_i \parallel N)$, $J = h(x \parallel ID_U)$. Next S checks if $h(T_i \parallel J)$?= C_1. If this holds, it means U_i is an authentic user, whereas the login request is rejected.
3. S acquires T_S, computes $C_2 = h(C_1 \oplus J \oplus T_S)$ & sends $\{C_2, T_S\}$ to U_i.
4. When receiving message from S, U_i checks time interval between T_S & $T^{''}$, where $T^{''}$ is the timestamp when mutual authentication message was received. If ($T^{''}$ - T_S) $\geq \Delta T$, then U_i rejects this message & terminates the session. Otherwise, U_i checks if $h(C_1 \oplus J \oplus T_S)$?= C_2. If this doesn't holds, U_i terminates session. Otherwise U_i & S share $SK = h(C_2 \oplus J)$.

Password Change Phase. In this phase, U_i can change his or her password.

1. U_i inserts SC into card-reader & inputs ID_i & pw_i.
2. SC computes $RPW^* = h(r \parallel pw_i)$ & $J^* = L \oplus RPW^*$. If J ?= J^* holds, then U_i is allowed to update password. Otherwise, this phase is terminated.
3. SC computes $L = J \oplus RPW \oplus RPW^* \oplus h(r \parallel pw^{'}_i)$ & replaces the old value of L with the new value. Now, the new password is updated.

Lost Smart Card Revocation Phase. U_i performs some steps to revoke SC.

1. S checks secret credentials's U_i, e.g. date of birth, identity card number.
2. S changes the value of N to revoke SC. In every case of stolen or lost of SC, N is increased by one. Later, U_i can re-register to S without changing ID_i.
3. S requests U_i to return to registration phase. Here, U_i is strongly recommended not to use any previous values for new registration, e.g. password & random value, otherwise anybody can impersonation U_i by using the same credentials previously saved in the lost or stolen SC.

2.2 Cryptanalysis of Muhammad Khurram Khan's Scheme

- Inability To Protect User's Anonymity: Khan et al claimed that only S can recover ID_i of U_i due to y used to hide the user's identity during transmission of login message. Hence, adversaries cannot identify the person trying to login into S. We see this explanation is not approriate because anyone being a valid user can know y. For example, another valid user captures $\{AID_i, T_i, d, C_1\}$. Then, he computes $ID_i = AID_i \oplus h(y \parallel T_i \parallel d)$
- Secret key forward secrecy: Khan et al claimed that even if the server's x & y happens to be compromised, an adversary cannot impersonate legitimate users by using the revealed keys, because he cannot compute AID_i & C_1 in the login message without knowledge of the user's ID_i, pw_i, r & ID_U. In this subsection, we will prove his claim is not true.

1. With y, adversary A can capture any login message & compute ID_i of any user by performing $ID_i = AID_i \oplus h(y \parallel T_i \parallel d)$.
2. A creates a new login message to impersonate U_i. Firstly, A picks T_A, random value d_A. Then, A computes $AID_i = ID_i \oplus h(y \parallel T_A \parallel d_A)$.
3. A assumes $N = 0$ & computes $J = h(x \parallel ID_U)$, where $ID_U = (ID_i \parallel N)$
4. A computes $C_A = h(T_A \parallel J)$ and sends $\{AID_i, T_A, d_A, C_A\}$ to S.
5. If everything is alright, A successfully impersonates U_i. Otherwise, A turns back to step 3 with increasing N by one & continues later steps.

– Denial of service attack: In Khan's scheme, passwords are not stored at the verification server. However, this scheme stores N value of U_i. If these values are modified by attackers, many users cannot login into server. For example, in authentication phase of Khan's scheme, S must compute $J = h(x \parallel ID_U)$, where $ID_U = (ID_i \parallel N)$. Then, S compares C_1 with $h(T_i \parallel J)$. So, if N is modified, C_1 is not equal to $h(T_i \parallel J)$ & S will reject login message.

3 Proposed Scheme

Our scheme is also divided into the five phases

3.1 Registration Phase

U_i submits ID_i, $h(r \parallel pw_i)$, where r is a nonce chosen by U_i. After receiving $\{ID_i, h(r \parallel pw_i)\}$ from user via a secure channel, S performs following steps.

1. Checking ID_i's existence. If it existed, S intimates U_i to choose another ID.
2. Generating a random value e & computing $J = h(x \parallel e)$, $P = h(J)$ & $L = J \oplus RPW$. Then S sends SC containing $\{L, e, P\}$ for U_i via a secure channel.
3. U_i receives SC & inputs r into it.

3.2 Login Phase

U_i inserts SC into card-reader, inputs ID_i & pw_i to login S. Then, SC performs:

1. Computing $RPW = h(r \parallel pw_i)$ & $J^* = L \oplus RPW$.
2. Checking whether $h(J^*)$?= P. If this holds, SC goes to next step. Otherwise, it terminates the session. Then, SC generates a random value R_U & computes $AID_i = ID_i \oplus R_U$, $C_1 = R_U \oplus J^*$ & $M_1 = h(ID_i \parallel J^* \parallel R_U)$.
3. Finally, SC sends $\{AID_i, C_1, M_1, e\}$ to S.

3.3 Mutual Authentication and Session Key Agreement Phase

S receives U_i's login message $\{AID_i, C_1, M_1, e\}$ and performs following steps.

1. S computes $J^{**} = h(x \parallel e)$, $R_U = C_1 \oplus J^{**}$ & $ID^*{}_i = AID_i \oplus R^*{}_U$.

2. S checks identity's validity. If everything isn't alright, S terminates the session, otherwise S checks whether $M_1 \stackrel{?}{=} h(ID^*_i \parallel J^{**} \parallel R^*_U)$. If this doesn't holds, S terminates the session, otherwise S generates a random value R_S & computes $C_2 = R_S \oplus J^{**}$, $M_2 = h(R_S \parallel J^{**} \parallel ID^*_i)$. Then S sends $\{M_2, C_2\}$ to user via a common channel.
3. After receiving $\{M_2, C_2\}$ from S. U_i computes $R^*_S = C_2 \oplus J^*$ & check if $M_2 \stackrel{?}{=} h(R^*_S \parallel J^* \parallel ID_i)$. If this does not hold, U_i terminates the session. Otherwise, U_i authenticates S successfully. U_i sends $M_3 = h(R_U \parallel R^*_S)$ to S & computes a session key $SK = h(R_U \parallel R^*_S \parallel J^* \parallel ID_i)$.
4. When receiving $\{M_3\}$ from U_i, S checks if $M_3 \stackrel{?}{=} h(R^*_U \parallel R_S)$. If this does not hold, S terminates the session. Otherwise, S authenticates U_i successfully. And S also computes $SK = h(R^*_U \parallel R_S \parallel J^{**} \parallel ID^*_i)$.

3.4 Password Update Phase

When U_i wants to change password pw_i. He can perform following steps:

1. Insert SC into card-reader, input ID_i, pw_i & choose a new password pw_{inew}.
2. SC computes $RPW = h(r \parallel pw_i)$ & $J^* = L \oplus RPW$. Then, SC checks if $h(J^*) \stackrel{?}{=} P$. If this doesn't holds, SC terminates the session. Otherwise, SC computes $L_{new} = J^* \oplus RPW_{new}$, where $RPW_{new} = h(r \parallel pw_{inew})$
3. Finally, SC replaces L_i with L_{inew}.

3.5 Lost Smart Card Revocation Phase

We also recommend user not to use any previous values for new SC, e.g. password and random value. Following are some steps to perform in this phase:

1. User inputs old ID_i, new password pw_i & new random value r. Then U_i sends {credentials, ID_i, RPW} to S via a secure channel.
2. After receiving this package of message, S checks ID_i's validity in database. If it does not exist, S terminates the session. Otherwise, S continues to check U_i's credentials. If everything is alright, S generates a new random value e.
3. S computes $J = h(x \parallel e)$, $L = J \oplus RPW$ & $P = h(J)$. Then S sends new SC containing $\{L, e, P\}$ to U_i. Finally, U_i updates random value r into SC.

4 Security and Efficiency Analysis

In this section, we analyze our scheme on two aspects: security and efficiency.

4.1 Security Analysis

1. Replay Attack: In our scheme's authentication phase, if adversary A captures $\{AID_i, C_1, M_1, e\}$, he still cannot re-send this package again. For example, A replays the package, but A cannot compute random value R_S from S because of lacking value J. So, our scheme resists this kind of attack.

2. User's Anonymity: In our scheme, if adversary A wants to know ID_i of user, A must know random value R_U. However, R_U is enrypted with the value J which is not be leaked. Therefore, our scheme can protect user's anonymity.
3. Stolen Verifier Attack: In our scheme, S does not store any user's information except identity, so our scheme can counteract this kind of attack. In our scheme, S generates a random value e for each user. Hence, when authenticating with S, U_i only needs to send e to S and S uses master key x to re-construct $h(x \parallel e)$ of that user. So, S doesn't need to keep U_i's password.
4. Denial of Service Attack: In Khan's scheme, author stores value N of each user. So, if all N in server's database are modified, all users cannot login to server in login phase. Unlike his scheme, our scheme does not store any user's information. Hence, our scheme is immune from this kind of attack.

Besides above attacks, our scheme also has security features similar to Khan's.

4.2 Efficiency Analysis

We reuse approach used in some previous schemes to analyze computational complexity. That is, we calculate the number of one-way hash function execution. Let T_h be the time to compute one-way hash function. Khan's scheme needs $2 \times T_h$ in registration phase, and $8 \times T_h$ in login and authentication phases. Our scheme needs $3 \times T_h$ in registration phase and $9 \times T_h$ in login and authentication phases. Besides, we see his scheme doesn't resist to denial of service attack and cannot protect user's anonymity. Our proposed scheme recovers two pitfalls successfully. However, we don't solve secret key forward secrecy yet.

Table 1. A comparison between our scheme & Khan's for withstanding various attacks

Kinds of Attacks \ Schemes	Khan's	Ours
Denial of service attack	No	Yes
User's anonymity	No	Yes
Insider attack	Yes	Yes
Stolen verification table attack	Yes	Yes
Secret key forward secrecy	No	No
Mutual authentication & SK exchange	Yes	Yes
Replay attack	Yes	Yes

5 Conclusions

In this paper, we review Khan's scheme. Although his scheme can withstand some attacks, we see that his scheme is still vulnerable to denial of service attack, secret key forward secrecy and cannot protect user's anonymity. Consequently, we propose an improved scheme to eliminate some problems existing.

References

[1] Lamport, L.: Password authentication with insecure communication. Communications of the ACM 24, 770–772 (1981)
[2] Lee, C.-C., Lin, T.-H., Chang, R.-X.: A secure dynamic id based remote user authentication scheme for multi-server environment using smart cards. Expert Syst. Appl. 38(11), 13 863–13 870 (2011)
[3] Shen, J.-J., Lin, C.-W., Hwang, M.-S.: A modified remote user authentication scheme using smart cards. IEEE Transactions on Consumer Electronics 49, 414–416 (2003)
[4] Das, M.L., Saxena, A., Gulati, V.P.: A dynamic id-based remote user authentication scheme. IEEE Transactions on Consumer Electronics 50(2), 629–631 (2004)
[5] Liao, I.-E., Lee, C.-C., Hwang, M.-S.: Security enhancement for a dynamic id-based remote user authentication scheme. IEEE Transactions on Consumer Electronics 50, 629–631 (2004)
[6] Yoon, E.-J., Yoo, K.-Y.: Improving the Dynamic ID-Based Remote Mutual Authentication Scheme. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM 2006 Workshops. LNCS, vol. 4277, pp. 499–507. Springer, Heidelberg (2006)
[7] Khana, M.K., Kimb, S.-K., Alghathbara, K.: Cryptanalysis and security enhancement of a more efficient & secure dynamic id-based remote user authentication scheme. Computer Communications 34(3), 305–309 (2010)

A Simplified Privacy Preserving Message Delivery Protocol in VDTNs*

Youngho Park, Chul Sur, and Kyung-Hyune Rhee

Department of IT Convergence and Application Engineering,
Pukyong National University, Busan, Republic of Korea
{pyhoya,kahlil,khrhee}@pknu.ac.kr

Abstract. In Vehicular Ad Hoc Networks (VANETs), because of high mobility of vehicles and frequent change of road segments, an end-to-end communication path between moving vehicles may not exist unfortunately. As a promising solution to this challenge, for non-realtime constrained VANET applications, store-carry-forward paradigm is considered to deliver a message to a remote destination vehicle effectively through a socialspot in city road environments. So, the behavior of VANET can be modeled as Delay Tolerant Networks, and known as Vehicular Delay Tolerant Networks (VDTNs). Therefore, in this paper, we propose a secure message delivery protocol for protecting receiver-location privacy in socialspot-based VDTNs since location privacy is one of critical security requirements in VANETs. To design a simplified protocol, we eliminate the use of pseudonym-based vehicle identification accompanied with a complex pseudonymous key management. Instead, we introduce an identity-hidden message indexing which enables a receiver vehicle to query a message whose destination is itself to the socialspot RSU without revealing its identity.

Keywords: VANET, VDTN, authentication, privacy preservation.

1 Introduction

Vehicular Ad Hoc Networks (VANETs) to support the Intelligent Transportation Systems and Telematics have recently become one of the promising wireless networking research areas. This trend is due to Dedicated Short Range Communications (DSRC) [14] and the GPS-based navigation system incorporating with digital map. Typically, in VANETs, each vehicle equips with an on-board unit (OBU) communication device, which allows Vehicle-to-Vehicle (V2V) communication with other vehicles as well as Vehicle-to-Infrastructure (V2I) communication with a road-side unit (RSU). With these deployments, the VANET enables useful applications in our daily lives ranging from safety related to non-safety

* This research was supported by Basic Science Research Program (No. 2012-0001331), and partially supported by Next-Generation Information Computing Development Program (No. 2011-0029927) through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 416–425, 2013.

related, such as not only cooperative driving and probing vehicle data for better driving environment but also infotainment services by vehicular communications.

However, an end-to-end communication path between moving vehicles may not exist unfortunately because vehicles are constantly moving with frequently changing road segments which, in turn, makes network connectivity change. As a promising solution to this challenge, for non-realtime constrained VANET applications, store-carry-forward paradigm is considered to deliver a message to a multi-hop destination vehicle effectively by utilizing socialspot tactic [7] in city road environments. Here, the socialspots are referred to the locations in a city road that many vehicles often visit such as intersections around famous shopping malls, restaurants, or cinemas. Hence, we can utilize an RSU installed in the socialspot as a relay node for packet forwarding in an opportunistic way [7][8][6]. So, the behavior of a multi-hop VANET communication can be modeled as a Delay Tolerant Network known as Vehicular Delay Tolerant Networks (VDTNs), and packet forwarding protocols using store-carry-forward manner have been proposed [13][4].

As VANETs have received a lot of attention, security issues, especially privacy of vehicles or drivers, have become one of the most concerns for the successful deployment of VANET. In the same vein, socialspot-based VDTN applications must protect vehicle's privacy even though the locations of socialspots for message dissemination are known. That is, a security mechanism should be able to make it difficult as far as possible for an adversary who knows the locations of socialspots to infer which vehicle receives a message from the RSU at each socialspot.

1.1 Related Work

In order to protect receiver-location privacy in VDTNs, Lu et al. proposed socialspot-tactic privacy-preserving data forwarding protocols in [7] and [8], respectively. Those protocols are on the basis of pseudonym-based vehicle identification for anonymous message delivery and receiver authentication[1]. Therefore, each vehicle has to have pre-loaded pseudonym-set for avoiding vehicle tracking by periodically changing its pseudonym on the road. However, they require complex pseudonym-based cryptographic key management depending on the number of pseudonyms pre-loaded, and all vehicles must know receiver vehicle's pseudonym to send a message to the receiver. What is worse, the protocol of [7] does not provide message source authentication so this protocol cannot guarantee the non-repudiation if a malicious vehicle sends a bogus message.

On the other hand, the authors [8] incorporated conditional privacy-preserving authentication based on group signature and universal re-encryption scheme with packet forwarding protocol for protecting vehicle's location privacy from packet analysis attack. However, when a receiver vehicle downloads a message it is required for the receiver vehicle to perform a complex mutual authentication

[1] When we say sender and receiver, they are end-to-end message source and destination vehicle in this paper, respectively.

process with RSU at the socialspot due to the much time consuming operation of group signature scheme.

1.2 Contribution and Organization

Based on the above observation, in this paper, we propose a socialspot-based secure message delivery protocol for preserving receiver-location privacy. The main design goal of this paper is to simplify the authentication process of a receiver vehicle to a socialspot RSU by eliminating the use of pseudonym-set. The contributions of this paper are summarized as follows.

- Instead of putting vehicles' pseudo-ID to identify a receiver vehicle in anonymous manner, we introduce an identity-hidden message indexing in order for a receiver vehicle to query the message bound for it to the socialspot RSU without revealing its identity.
- We establish a unidirectionally authenticated secure message delivery channel from a sender to a receiver for VDTNs in which an interactive message exchange is not always possible because of no simultaneous end-to-end connection.
- To simplify the authentication process between a receiver vehicle and a socialspot RSU without presenting receiver's identity-related information, we make the receiver vehicle be implicitly authenticated to the RSU by proving knowledge of the shared secret key with the sender.

To design the proposed protocol, we make use of ID-based non-interactive key agreement scheme [12][3] (but the IDs of vehicles are not included in message delivery protocol) to establish a secure channel between sender and receiver vehicles, and cryptographic hash function to generate an identity-hidden message index while binding a specific receiver vehicle at a socialspot is possible. The remainder of this paper is organized as follows. In Section 2, we introduce our system model and security goals considered in this paper. We present the proposed protocol and provide security analysis in Section 3 and Section 4, respectively. Finally, we conclude in Section 5.

2 System Model

We consider a VDTN environment which consists of vehicles equipping with OBUs, RSUs installed in socialspots and Trusted Authority(TA) for security management as shown in Fig. 1, respectively.

- TA is in charge of issuing ID-based private keys to the registered vehicles and RSUs, and provides public system parameters for security protocol.
- Socialspots $\mathcal{SS} = \{ss_1, ..., ss_l\}$ are referred to as roads or intersections around where many vehicles will visit, for example, famous shopping malls, movie theaters, and such like. At each $ss_j \in \mathcal{SS}$, a huge-storage possessing RSU_j subordinated by the TA is installed so that RSU_j can temporarily store some messages to be forwarded to the receiver vehicles passing through the ss_j.

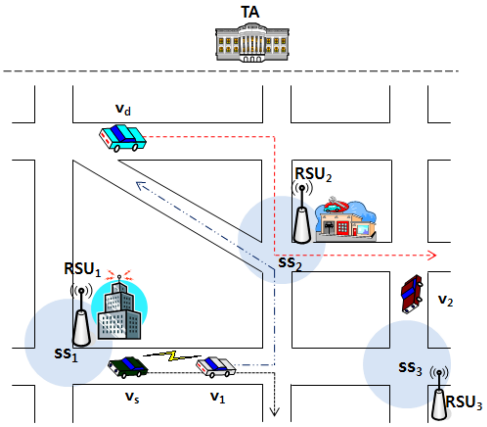


Fig. 1. System model for socialspot-based VDTN

– Each vehicle $v_i \in \mathcal{V} = \{v_1, ..., v_n\}$ registered to the system equips with OBU for V2V and V2I communications and cooperates with each other to deliver a message for a socialspot in store-carry-forward manner.

In those settings, message forwarding strategy from a sender vehicle to a destination socialspot can be divided into the following two methods.

– Direct carry and forward : If the sender vehicle passes the socialspot, the sender will carry the message and then forward it when it arrives on the socialspot.
– V2V forward and carry : Some vehicles driving toward the socialspot will cooperate for store-carry-forward message delivery when the sender vehicle does not pass the socialspot.

As an example scenario, suppose that v_s wants to send a message msg to v_d which will visit socialspot ss_2 later in Fig. 1.

1. At time t_1, v_s asks v_1 which drives toward the ss_2 for forwarding the msg.
2. v_1 carries the msg and arrives on the socialspot ss_2 at time $t_2 (t_2 > t_1)$, then forwards the msg to the RSU_2.
3. When v_d passes the ss_2 at time $t_3 (t_3 > t_2)$ while RSU_2 stores the msg, v_d requests msg bound for it then RSU_2 provides v_d with msg.

In such a VDTN scenario, we consider the following security goals to design a secure message delivery protocol against a global passive adversary $\mathcal{A}$. The adversary $\mathcal{A}$ can overhear V2V and V2I communications, but cannot compromise any vehicle (or RSU) and access the internal information of them. Thus, $\mathcal{A}$ tries to identify vehicles or to trace the location of a vehicle by packet analysis.

– *Anonymous Channel* : An adversary $\mathcal{A}$ cannot identify the message sender and receiver from eavesdropping on the message delivery protocol.

- *Authentication* : Only a valid receiver vehicle specified by a sender can retrieve the message whose destination is itself by authenticating itself to the RSU at a socialspot.
- *Receiver Privacy* : Even though the location of a socialspot is known, it is hard for an adversary $\mathcal{A}$ to infer which vehicles retrieved messages at the socialspot.

3 Proposed Protocol

The proposed protocol consists of *setup*, *message constitution*, *message forwarding*, and *message retrieving* phases. Table 1 shows the notations and descriptions used in the proposed protocol.

Table 1. Notations and descriptions

notation	description
SK_i	ID-based private key of an entity i
k_{ij}	shared secret key between i and j
T	valid time period of a message
$Enc_k(\cdot)$	encryption under key k
$Dec_k(\cdot)$	decryption under key k
$Sig_{SK_i}(\cdot)$	ID-based signature under signing key SK_i
$Vrf_i(\cdot)$	ID-based signature verification for a given ID i
$h(\cdot)$	cryptographic hash function
$MAC_k(\cdot)$	message authentication code under key k

3.1 Setup

Let $\mathbb{G}_1$ and $\mathbb{G}_2$ be the bilinear map groups with a prime order q, and P be a generator of $\mathbb{G}_1$ [1], respectively. In the setup phase, the TA configures system parameters for bilinear map and issues ID-based private keys to the registered RSUs and vehicles as following steps.

1. TA sets a random number $s \in \mathbb{Z}_q^*$ as its master secret key, computes $P_0 = sP$, and configures public system parameters $param$=$\{\mathbb{G}_1, \mathbb{G}_2, q, \hat{e}, P, P_0, H_1, H_2\}$, where $\hat{e} : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$ is a bilinear map, $H_1 : \{0,1\}^* \rightarrow \mathbb{G}_1$ and $H_2 : \{0,1\}^* \rightarrow \mathbb{Z}_q^*$ are cryptographic hash functions, respectively.
2. For each $v_i \in \mathcal{V}$ and each RSU_j at $ss_j \in \mathcal{SS}$, TA issues ID-based private keys $SK_{v_i} = sH_1(v_i)$ for v_i and $SK_j = sH_1(ss_j)$ for RSU_j, respectively.

3.2 Message Constitution

When a vehicle v_s wants to send a message msg to a receiver vehicle v_d which will pass a socialspot ss_j sometime, v_s executes the following steps to make a packed message.

1. v_s chooses a random number $r \in \mathbb{Z}_q^*$ and computes $P_s = rH_1(v_s)$, $k_{sd} = \hat{e}(rSK_{v_s}, H_1(v_d))$, $k_{sj} = \hat{e}(rSK_{v_s}, H_1(ss_j))$, $w = H_2(k_{sd}|T)$, and $W = w^{-1}P$, where k_{sd} and k_{sj} are non-interactively shared keys with v_d and with RSU_j, respectively.
2. Then, v_s constitutes a packed message M to be forwarded to the destination socialspot ss_j as follows:
   $M = \{ss_j, I, P_s, W, C|\sigma, chk\}$
   - $I = h(v_d, ss_j, T)$
   - $C = Enc_{k_{sd}}(v_s|v_d|T|msg)$
   - $\sigma = Sig_{SK_{v_s}}(v_s|v_d|T|msg)$
   - $chk = MAC_{k_{sj}}(ss_j, I, P_s, W, C|\sigma)$

   where σ is sender v_s's ID-based signature of [2], and chk is a message authentication code for integrity check by RSU_j.

In step 2, the identity-hidden message index I implies the meaning of a receiver vehicle v_d at a socialspot ss_j, and will be used for a receiver vehicle to query a message for it in the message retrieving phase.

3.3 Message Forwarding

Once the message M is packed, M can be delivered to a destination socialspot ss_j by using the forwarding strategy described in Section 2. At this phase, we assume a packet forwarding protocol with store-carry-forward manner, such as VADD [13] and TBD [4]. Note that the main goal of this paper is to protect receiver's privacy from an adversary, we do not consider compromising of vehicles and message forgery attack by an active adversary during the message forwarding.

When the message M ultimately reaches the RSU_j at ss_j, RSU_j first computes shared key of v_s as $k_{sj} = \hat{e}(P_s, SK_{ss_j})$ from P_s in M. Then, RSU_j verifies $chk = MAC_{k_{sj}}(ss_j, I, P_s, W, C|\sigma)$ under the key k_{sj}. If chk is valid, RSU_j stores $\{I, P_s, W, C|\sigma\}$ while a receiver vehicle related to the message index I requests the message as passing by it.

3.4 Message Retrieving

Fig. 2 shows the message retrieving protocol of a receiver vehicle at a socialspot. When a vehicle v_d goes by a socialspot ss_j on its way driving, v_d can get a message M whose destination is itself as follows.

1. v_d, as expecting a message for it on RSU_j's storage, generates its message index at ss_j as $I = h(v_d, ss_j, T)$, then queries I to RSU_j.
2. RSU_j searches its storage for the message corresponding to I. If the message is found, RSU_j sends P_s of matching index I to v_d as a challenge for authentication.
3. Upon receiving P_s, v_d computes the secret key $k_{sd} = \hat{e}(P_s, SK_{v_d})$ shared with a sender and $w = H_2(k_{sd}|T)$, then gives $\widetilde{W} = wP$ to the RSU_j as a proof of knowledge of the shared key.

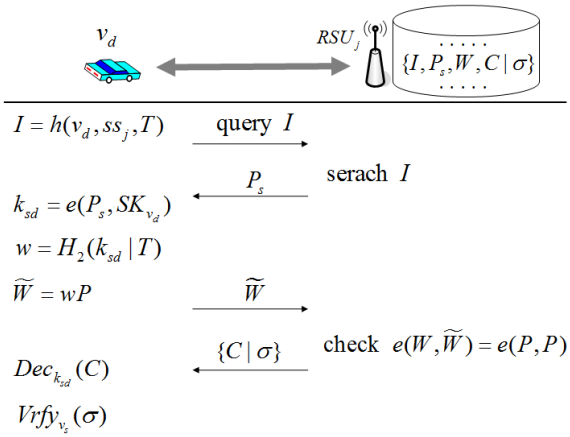


Fig. 2. Message retrieving protocol of a receiver vehicle at a socialspot

4. With W sent from a sender v_s and $\widetilde{W}$ from v_d, RSU_j verifies $\hat{e}(W, \widetilde{W}) = \hat{e}(P, P)$ to check the proof of knowledge. If the verification holds, RSU_j authenticates v_d as a valid receiver specified by the sender, then provides $\{C|\sigma\}$ to v_d.
5. v_d recovers $\{v_s|v_d|T|msg\}$ by computing $Dec_{k_{sd}}(C)$, and finally completes the message retrieving protocol after verifying the signature σ as $Vrf_{v_s}(\sigma)$.

4 Analysis

In this section, we discuss the security of the proposed protocol. The security of the proposed protocol entirely depends on the non-interactive key agreement scheme and cryptographic hash function. We will focus on how the proposed protocol can fulfil our security goals under our adversary model.

4.1 Anonymous Channel

In the proposed protocol, the delivered message content $\{v_s|v_d|T|msg\}$ from a sender v_s to a receiver v_d is encrypted under non-interactively shared key k_{sd}, i.e., $C = Enc_{k_{sd}}(\{v_s|v_d|T|msg\})$. Hence, when we assume the secrecy of non-interactive key agreement scheme [3], it is difficult for an adversary $\mathcal{A}$ to identify sender and receiver from eavesdropping on the message transmission. Even if $\mathcal{A}$ can know that the destination of the message is a socialspot ss_j, $\mathcal{A}$ cannot capture the identities of vehicles which retrieve messages through the socialspot RSU_j because no vehicle identity is presented to the RSU_j. Therefore, the proposed protocol can guarantee the anonymity of message transmission.

In addition, Kate et al. [5] presented that they could construct an onion routing for anonymity network on the basis of non-interactive key agreement scheme. If we encrypt the packed message M again under key k_{sj} instead of

$MAC_{k_{sj}}$ in Message Constitution phase, the path $v_s \rightarrow ... \rightarrow RSU_j \rightarrow v_d$ can be regarded as an onion path based on Kate et al.'s observation.

4.2 Authentication

In order to obtain a message temporarily stored in a RSU_j in Message Retrieving phase, a receiver vehicle must be authenticated to the RSU_j which checks if the requesting vehicle is the designated receiver by a sender vehicle. In our protocol, for a vehicle v_d to be authenticated as a valid receiver, v_d should present the proof of knowledge $\widetilde{W} = H_2(k_{sd}|T)P$ for the secret key k_{sd} shared with a sender v_s. The consistency of the keys $\hat{e}(rSK_{v_s}, H_1(v_d))$ generated by v_s and $\hat{e}(P_s, SK_{v_d})$ by v_d can be proven as $\hat{e}(rSK_{v_s}, H_1(v_d)) = \hat{e}(rH_1(v_s), sH_1(v_d)) = \hat{e}(P_s, SK_{v_d})$.

Only if the verification of $\hat{e}(W, \widetilde{W}) = \hat{e}(P, P)$ holds, RSU_j will send $\{C|\sigma\}$ to v_d as regarding v_d is the receiver who can agree with the message sender. Then, v_d can recover original message $\{v_s|v_d|T|msg\}$ by decrypting C, and authenticates the sender v_s as verifying v_s's signature σ.

4.3 Receiver Privacy

As mentioned before, the proposed protocol does not put vehicle's identity for message transmission nor receiver's identity is given to the RSU_j at a socialspot ss_j in message retrieving phase. Instead, a receiver v_d can be bound by identity-hidden message index $I = h(v_d, ss_j, T)$ which is the result of cryptographic one-way hash function. Therefore, it is hard for an adversary $\mathcal{A}$ to decide which vehicle receives a message from I at the socialspot even though the location of the socialspot is publicly known.

Moreover, we can generate a different message index I' ($\neq I$) for different time period or different socialspot, i.e., $I' = h(v_d, ss_j, T')$ for $T' \neq T$, due to the functionality of cryptographic hash function. Hence, the proposed protocol can guarantee the unlinkability for a receiver vehicle because it is infeasible for $\mathcal{A}$ to distinguish that the given indexes I' and I are linked to the same receiver.

However, one feasible attack for $\mathcal{A}$ is to prepare possible message index set $\mathcal{I}_S$ from arbitrarily chosen vehicles identities $\mathcal{V}_A = \{v_1, ..., v_m\}$ by $\mathcal{A}$ for a given time period T, and check if an $I \in \mathcal{I}_S$ occurs at the socialspot ss_j or not. If it occurs, then $\mathcal{A}$ can decide the matching identity $v_i \in \mathcal{V}_A$ such that $I = h(v_i, ss_j, T)$. For this scenario, let $Pr\{k\}$ be the probability that k among N_V vehicles passing through the socialspot ss_j for the given time period T are found by the index matching attack. Suppose that N_R is total number of registered vehicles and N_C is the number of elements in $\mathcal{I}_S$. The probability $Pr\{k\}$ can be represented as follow distribution:

$$Pr\{X = k\} = \frac{\binom{N_C}{k}\binom{N_R - N_C}{N_V - k}}{\binom{N_R}{N_V}}, \quad k \geq 1$$

As a result, the probability that a target vehicle v_d can be linked to $\mathcal{I}_S$ is $Pr\{k = 1\}$. Fig. 3 shows such link probability under chosen message index matching attack by $\mathcal{A}$ assuming $N_R = 10,000$ for evaluation. From this result, we can figure out that the link probability decreases as the number of vehicles N_V passing through a socialspot increases. Therefore, we can conclude that putting a special area where many vehicles frequently visit in city road environments as a socialspot is helpful for privacy preservation for secure message delivery in VDTNs.

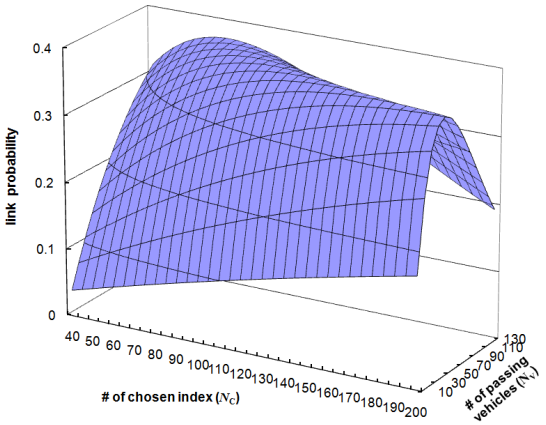


Fig. 3. Link probability for $k = 1$ under chosen message index matching by $\mathcal{A}$

5 Conclusion

In this paper, we proposed a secure message delivery protocol with the help of socialspots in Vehicular Delay Tolerant Networks to provide anonymous message transmission and vehicles privacy preservation assuming a global passive adversary. To design a simplified protocol, we eliminated the pseudonym-based receiver vehicle identification accompanied with a complex pseudonymous key management. Instead, we made use of identity-hidden message indexing for a receiver vehicle to prevent vehicle's identity from being disclosed or linked by an adversary, and proof of knowledge for non-interactively shared key between sender and receiver to authenticate the receiver implicitly by a socialspot RSU. In addition, we showed that it is infeasible for an adversary to link a specific vehicle to a message index at a socialspot.

References

1. Boneh, D., Franklin, M.: Identity-based Encryption from the Weil Pairing. SIAM Journal of Computing 32(3), 586–615 (2003)
2. Cha, J., Cheon, J.: Identity-Based Signature from Gap Diffie-Hellman Groups. In: Desmedt, Y.G. (ed.) PKC 2003. LNCS, vol. 2567, pp. 18–30. Springer, Heidelberg (2002)

3. Dupont, R., Enge, A.: Provably secure non-interactive key distribution based on pairings. Discrete Applied Mathematics 154(2), 270–276 (2006)
4. Jeong, J., Guo, S., Gu, Y., He, T., Du, D.H.C.: Trajectory-based data forwarding for light-traffic Vehicular Ad Hoc Networks. IEEE Transaction on Parallel and Distributed Systems 22(5), 743–757 (2011)
5. Kate, A., Zaverucha, G.M., Goldberg, I.: Pairing-based onion routing with improved forward secrecy. Journal of ACM Transactions on Information and System Security, TISSEC 13(4), Article No. 29 (2010)
6. Lin, X., Lu, R., Liang, X., Shen, X.: STAP: A social-tier-assisted packet forwarding protocol for achieving receiver-location privacy preservation in VANETs. In: Proceedings of the IEEE INFOCOM, pp. 2147–2155 (2011)
7. Lu, R., Lin, X., Liang, X. (Sherman) Shen, X.: Sacrificing the plum tree for the peach tree: A socialspot tactic for protecting receiver-location privacy in VANET. In: Proceedings of the IEEE Global Telecommunications Conference, GLOBECOM, pp. 1–5 (2010)
8. Lu, R., Lin, X., Shen, X.: SPRING: A social-based privacy-preserving packet forwarding protocol for Vehicular Delay Tolerant Networks. In: Proceedings of the IEEE INFOCOM, pp. 632–640 (2010)
9. Lu, R., Lin, X., Zhu, H., Ho, P.-H., Shen, X.: ECPP: Efficient conditional privacy preservation protocol for secure vehicular communications. In: Proceedings of the IEEE INFOCOM, pp. 1229–1237 (2008)
10. Nzouonta, J., Rajgure, N., Wang, G., Borcea, C.: VANET routing on city roads using real-time vehicular traffic information. IEEE Transactions on Vehicular Technology 58(7), 3609–3626 (2009)
11. Raya, M., Hubaux, J.-P.: Securing vehicular ad hoc networks. Journal of Computer Security 15(1), 39–68 (2007)
12. Sakai, R., Ohgishi, K., Kasahara, M.: Cryptosystems based on pairing. In: Symposium on Cryptography and Information Security, SCIS 2000 (2000)
13. Zhao, J., Cao, G.: VADD: Vehicle-assisted data delivery in Vehicular Ad Hoc Networks. In: Proceedings of the IEEE INFOCOM, pp. 1–12 (2006)
14. Dedicated Short Range Communications (DSRC), `http://www.leearmstrong.com/dsrc/dsrchomeset.htm`

Confidentiality-Preserving Query Execution of Fragmented Outsourced Data

Anis Bkakria[1], Frédéric Cuppens[1], Nora Cuppens-Boulahia[1], and José M. Fernandez [2]

[1] Télécom Bretagne
[2] École Polytechnique de Montréal
{anis.bkakria,frederic.cuppens,nora.cuppens}@telecom-bretagne.eu, jose.fernandez@polymtl.ca

Abstract. Ensuring confidentiality of outsourced data continues to be an area of active research in the field of privacy protection. Almost all existing privacy-preserving approaches to address this problem rely on heavyweight cryptographic techniques with a large computational overhead that makes inefficient on large databases. In this paper, we address this problem by improving on an existing approach based on a combination of fragmentation and encryption. We present a method for optimizing and executing queries over distributed fragments stored in different Cloud storage service providers. We then extend this approach by presenting a Private Information Retrieval (PIR) based query technique to enforce data confidentiality under a collaborative Cloud storage service providers model.

Keywords: Data confidentiality, Privacy-preserving, Data fragmentation, Data outsourcing.

1 Introduction

In the last few years, database outsourcing has become an important tool in IT management, as it offers several advantages to the client companies, especially for small ones with limited IT budget. In most models of database outsourcing, data storage and management (e.g. data backup and recovery) are completely operated by external service providers that take advantages of economies of scale to reduce the cost of maintaining computing infrastructure and data-rich applications in comparison with the cost of in-house data management.

Nonetheless, outsourcing gives rise to major security issues due to the usually sensitive nature of information saved in databases. Storing such sensitive information with external service providers requires them to be trusted and that they should protect data against external threats. Even if these service providers protect data adequately and process queries accurately, they may be curious about the stored data and may attempt to profit by disclosing this sensitive information to third parties. This raises the question of whether it is possible to protect confidentiality and privacy of the outsourced data. A recent study [10]

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 426–440, 2013.

by the Ponemon Institute indicates that since 2005 more than 250 million customer records containing sensitive information have been lost or stolen. The same study reports the average organizational cost of data leakage between 2005 and 2009, to be about $6.6 million. Hence, there seems to be a clear financial motivation to protect data privacy and confidentiality even when records are leaked.

Related Work. One approach to protect confidentiality and privacy of the outsourced data is based on encrypting all tuples in outsourced databases [8, 7]. With this approach, one crucial question is that of how to efficiently execute queries. Clearly, when using deterministic encryption techniques equality-match queries are simple to evaluate. However, range queries and aggregations become more difficult to perform as we must decrypt all records to evaluate this kind of queries, which makes query execution on the outsourced encrypted data much more difficult. Therefore, our main focus in this paper is about preserving both privacy and confidentiality of outsourced database while ensuring a secure way for querying outsourced data. One promising method to meet this requirement is to use data fragmentation. Basically, data fragmentation procedures are not particularly designed for preserving data security, they are aimed to improve data manipulation process, optimize storage, and facilitate data distribution. Nevertheless, in the last few years two significant alternatives have been proposed. The first one [9, 3] relies on data fragmentation alone to protect confidentiality. The distribution model used in this approach is composed mainly of two domains: a trusted local domain from which the data originates, and a honest but curious domain in which the data will be distributed. Because of its trustworthiness, the local domain is used to store fragments that contain highly sensitive data without the need to encrypt them. This is not so efficient in that it forces data owners to always protect and manage fragments containing highly sensitive data. A more promising alternative [4, 6] attempts to combine data fragmentation together with encryption. In this proposal, the main idea is to partition the data to be externalized across two or more independent service providers, and furthermore to encrypt all information which can not be secured by fragmentation (e.g. employees' bank account numbers of a company). While they do constitute an interesting way to ensure confidentiality of outsourced database, these approaches have the major limitation that it assumes that data to be outsourced is represented within a single relation schema (or table). Clearly, this assumption is too strong and seldom satisfied in the real environments, as generally, relational databases are normalized to minimize redundancy and dependency by dividing large tables into smaller (and less redundant) tables and defining relationships between them.

In this paper, we strive to protect the confidentiality and the privacy of sensitive outsourced database using both encryption and fragmentation, combining the best features of both approaches. Furthermore, we aim to overcome the previously mentioned limitations of [4, 6] by presenting an approach which is able to deal efficiently with multi-relation normalized databases. In a distributed

environment, the problems encountered in one-relation[1] databases take on additional complexity when working with multi-relation normalized databases as it gives rise to new problems such as protecting the relatioships between relational schemas and defining a secure and efficient technique allowing authorized users to query these sensitive relationships. We will show how to protect data unlinkability of different fragments of the original database by protecting user query privacy using a practical Private Information Retrieval (PIR) technique. Unlinkability of two items of interest (e.g., records stored into different fragments) means that within the system, from an adversary point of view, these items of interest are no more and no less related. In our approach, a relation containing sensitive information will be fragmented into two or more fragments. Unlinkability of fragments means that despite the fact that an adversary has knowledge about the fragments of a relation, it remains unable to link records from different fragments.

The remainder of the paper is organized as follows. Section 2 illustrates through an example the problem and need for an approach like ours. In Section 3, we detail the threat model, security model, and assumptions. Section 4 describes our approach to enforce privacy and confidentiality of outsourced data. Section 5 presents the query optimization and execution model. In Section 6, we present a PIR-based technique to achieve query privacy and enforce data confidentiality under a collaborative Cloud storage service providers model. Finally, we conclude the paper in section 7.

2 Motivating Example

Consider a relational hospital database D with relations: **Patient**(Id, Name,ZIP, Illness, Id_Doctor *) and **Doctor**(Id, Name, Specialty) where Id_Doctor is a foreign key referencing the column Id of the relation $Doctor$. Let us assume that the hospital is going to outsource the database to a third party. Sensitive information contained in D should be protected. If we look carefully, we can consider that the list of patients and their attributes $(Id, Name, Zip)$ are not sensitive, and also that the list of illnesses could be made public. Nevertheless, the relationship between these two lists (list of patients and list of illnesses) should be protected. Therefore if we can find a way (e.g. vertical fragmentation [11]) to break relationships between patients and their respective illnesses, there is no need to encrypt all records of the $Patient$ relation. On the other hand, the relationship between a patient and his doctor should be protected. Since the list of doctors itself is not sensitive, the simplest way to protect the relationship between the two relations $Patient$ and $Doctor$ consists in encrypting the foreign key Id_Doctor. Actually, we can either encrypt the foreign key Id_Doctor of the relation $Patient$ or the primary key Id of the relation $Doctor$, because in the two cases, relationship between relations $Patient$ and $Doctor$ will be protected. However, encrypting the foreign key seems to be more beneficial as a foreign key references only one relation (only the relationship between the two relations is

[1] Databases composed from a single relation shema.

protected) while a primary key can be referenced by many relations. Therefore, if we encrypt the primary key, we will protect all relationships between the relation containing the primary key and other relations referencing the encrypted primary key. Thus, when the security requirement specifies that only the a relationship between data is sensitive, our apporach is more appropriate than the one based on full encryption.

3 Threat Model and Security Assumptions

Our approach is based on a typical client-server architecture, where servers are managed by different service providers. These service providers are considered "honest-but-curious", in agreement with most related work [3, 4, 7, 9]. Service providers are assumed to be "honest" in that they do not manipulate outsourced data in order to respond incorrectly to user queries. In other words, we suppose that responses to user queries received from these service providers are always accurate. In the first part of this paper, we will assume that service providers are "curious" in that they will try to infer and analyze outsourced data, and will also actively monitor all received user queries and try to derive as much information as possible from these queries. In the second part of the paper, we further assume that service providers can collude and cooperate together to link outsourced data. The client part of this architecture is assumed to be trustworthy and all interactions between the user and the client are secured. Protecting the client part against external attacks is beyond the scope of this article.

4 Confidentiality Using Fragmentation and Encryption

Our approach extends in several ways the vertical fragmentation-based approach described in [4, 6]. This approach considers that all data is stored in a single relation, while in our approach data can be stored in several relations, which is the case in typical database environments. In our approach, we consider that databases to be externalized are normalized so that two relations can be only associated together through a primary key/foreign key relationship. For this purpose, we introduce a new type of confidentiality constraint for fragmentation, the *inter-table fragmentation constraint*. The aim of this new fragmentation constraint is to protect the relationship between relations. This section first presents the different kinds of confidentiality constraints used to achieve our goals of protecting the confidentiality by encryption and fragmentation, and second formalises the concept of fragmentation in our approach which extends ideas presented in [4–6].

Definition 1 (Confidentiality Constraint). Consider that data to be secured are represented with a relational database D, which is composed of a list of relational schemas $R = (R_1, \ldots, R_n)$, with each of these relational schemas R_i containing a list of attributes $A_{R_i} = (a_{1,i}, a_{2,i}, \ldots)$. A confidentiality constraint over D can be one of the following:

Singleton Constraint (SC). It is represented as a singleton set $SC_{R_i} = \{a_{j,i}\}$ over the relation R_i. This kind of confidentiality constraint means that the attribute $a_{j,i}$ of the relational schema R_i is sensitive and must be protected, typically by applying encryption.

Association Constraint (AC). This kind of confidentiality constraint is represented as a subset of attributes $AC_{R_i} = \{a_{1,i}, \ldots, a_{j,i}\}$ over the relational schema R_i. Semantically, it means that the relationship between attributes of the subset AC_{R_i} is sensitive and must be protected.

Inter-table Constraint (IC). It is represented as a couple of relational schemas $IC = \{R_i, R_j\}$ of the relational database D. Relations R_i and R_j should be associated through a primary key/foreign key relationship. The use of this kind of confidentiality constraint ensures protection of the primary key/foreign key relationship between the two relational schemas concerned with the inter-table constraint IC.

Note that protecting the relationship between two tables relies on protecting the primary key/foreign key relationship and storing involved relations separately. The association constraint can also be addressed through encryption (encrypt at least one of attributes involved in the constraint), but clearly this will increase the number of encrypted attributes and make interrogation of the database more complicated. A more adapted way to resolve this kind of confidentiality constraint was proposed in [4], which is based on splitting involved attributes in a manner that their relationships cannot be reconstructed.

In the case of an inter-table confidentiality constraint, protecting the foreign key using encryption is the simplest way to secure the relationship between the two relational schemas. However encrypting only the foreign key is not enough to keep the relationship between relational schemas secure, as service provider may be able to link records in two relational schemas by observing and analyzing user queries over these relational schemas. To overcome this problem, the two relational schemas in question should be split into different fragments, and each of these fragments should be distributed to a different Cloud storage provider. An interesting approach for modeling constraints and resolving the data fragmentation problem was proposed in [6], that efficiently computes data fragments that satisfy the confidentiality constraints. It is based on Boolean formulas and Ordered Binary Decision Diagrams (OBDD) and uses only attribute-based confidentiality constraint (Singleton Constraints and Association Constraints). However, it cannot deal as-is with Inter-table Constraints. In order to use this approach, we define a way to reformulate Inter-table Constraint as a set of Singleton Constraints and Association Constraints. We explain this transformation in the definition and theorem below.

Definition 2 (Inter-table Constraint transformation). Consider a relational database with two relations $R_1(\underline{a_1}, \ldots, a_n)$ and $R_2(\underline{b_1}, \ldots, b_m*)$. Let us assume that R_1 and R_2 are related through a foreign key/primary key relationship in which the foreign key b_m of the relation R_2 references the primary key a_1 of relation R_1. We assume that R_1 and R_2 contain respectively p and q records,

with $p > 1$ and $q > 1$. An Inter-table Constraint $c = \{R_1, R_2\}$ over relations R_1 and R_2 states that the relationship between these two relations must be protected by encrypting the foreign key b_m and by storing R_1 and R_2 in two different fragments. Therefore, the constraint c can be written as follows:

i A singleton constraint $SC = \{b_m\}$ to state that the value of b_m should be protected.
ii A list of $(m \times n)$ association constraints $AC = \{(a_i, b_j) | i \in [1, n],\ j \in [1, m]\}$.

We propose the notion of a correct transformation of Inter-table constraints. A transformation of an Inter-table constraint c to a set of confidentiality constraints C is correct if the satisfaction of C enforce the protection of the unlinkability between records of the two relations involved in c. The following Theorem formalizes this concept.

Theorem 1 (Transformation correctness). Given a relational database $\mathcal{D}$ composed from two relational schemas $R_1(\underline{a_1}, \ldots, a_n)$ and $R_2(\underline{b_1}, \ldots, b_m*)$ related through relationship between the foreign key b_m of R_2 and the primary key a_1 of R_1. Let $c = \{R_1, R_2\}$ be an Inter-table constraint, the set of constraints C be the result of the transformation of c, and $\mathcal{F} = \{F_1, \ldots F_q\}$ be a fragmentation of $\mathcal{D}$ that satisfies C. The Inter-table constraint c is correctly transformed into a set of constraints C if all the following conditions hold :

1. b_m does not appear in clear in any fragment of $\mathcal{F}$.
2. $\forall\ AC_{i,j} = \{a_i, b_j\} \in C, i \in [1, n], j \in [1, m],\quad if\quad a_i \in F_k\quad and\quad b_j \in F_l\quad then\ k \neq l$

The main advantage of the Inter-table constraint is that it allows treatment of multi-table relational databases. In addition, it gives a simple way to formulate confidentiality constraints between relations. As we have seen in Item (i) of Definition 2, the attribute b_m (foreign key of the relation R_2) should be encrypted. However, to be able to query data and construct relationship between relations, the chosen encryption algorithm must be deterministic [1] in order to preserve uniqueness and allow the construction of relationship between relations (e.g. through JOIN queries). As is known, in normalized multi-relation databases, three types of relationship between relations exist: (1) one-to-one, (2) one-to-many and (3) many-to-many relationships. Inter-table constraints over relations associated using (1) or (2) can be simply transformed as shown in Definition 2, while others associated using (3) need a pre-transformation step before applying the transformation of Definition 2, as they are normally linked through a third relation known as a linking table. The pre-transformation steps is described in the example below.

Definition 3. Fragmentation [5]
Let us consider a relational database D with relations $R_1, \ldots, R_n$ and A the list of all attributes contained in different relations. Given A_f the list of attributes to be fragmented, the result of the fragmentation is a list of fragments $F = \{F_1, \ldots, F_m\}$ where each of these fragments satisfies:

i $\forall F_i \in F, i \in [1..m],\ F_i \subseteq A_f$.
ii $\forall a \in A_f, \exists F_i \in F : a \in F_i$.
iii $\forall F_i, F_j \in F, i \neq j : F_i \cap F_j = \emptyset$.

Note that the list of attributes to be fragmented A_f contains all attributes in A, except those concerned with Singleton Constraints (attributes to be encrypted). Condition (i) guarantees that only attributes in A_f are concerned by the fragmentation, condition (ii) ensures that any attribute in A_f appears in clear at least in one fragment and condition (iii) guarantees unlinkability between different fragments.

Logically, to be able to get information about the original database, we should be able to reconstruct original database from fragments. So after defining the fragmentation process, we shall define a mechanism to combine fragmentation and encryption. More clearly, a mechanism to integrate attributes involved in the Singleton Constraints (attributes to be encrypted) in the suitable fragment. These encrypted attributes allow only authorized users (users who know the encryption key) to construct the sensitive relationships. Based on the definition of $Physical fragment$ proposed in [4], we define our mechanism called $Secure fragment$ to combine fragmentation and encryption.

Definition 4. Secure Fragment: Let D be a relational database with a list of relations $R = \{R_1(a_{1,1}, \ldots, a_{j,1}), \ldots, R_n(a_{1,n}, \ldots, a_{k,n})\}$, $F = \{F_1, \ldots, F_m\}$ a fragmentation of D and A_f be the list of fragmented attributes. Each fragment $F_i \in F$ is composed of a subset of attributes $A_i \subseteq A_f$. Each A_i is composed of a subset of attributes of one or more relations $R_j \in R$. We denote by R_{F_i} the list of relations in R which a subset of their attributes belongs to the fragment $F_i \in F$. The secure fragment of F_i is represented by a set of relations schema $R^e_{F_i}$ in which each relation is represented as follows $R^e_j(\underline{salt}, enc, a_1, \ldots, a_k)$ where $\{a_1, \ldots, a_k\} \subset A_i \cap R_j$ and enc is the encryption of all attributes of R_j that do not belong to $\{a_1, \ldots, a_k\}$ (all attributes of R_j involved in a singleton constraint except those concerned by a singleton constraint over the foreign key), combined before encryption in a binary XOR with the salt. All foreign key attributes which are involved in singleton constraints are encrypted using a deterministic encryption algorithm (e.g., AES) to ensure their indistinguishability. The Algorithm 1 shows the construction of secure fragments.The main reason for reporting all original attributes (except foreign keys involved in the Singleton constraints) in an encrypted form for each relation in a fragment, is to guarantee that a query Q over the original relation R_j can be executed by querying a single fragment (which contains R^e_j) while preserving confidentiality of sensitive relationships, so we do not need to reconstruct the original relation R_j to perform the query Q. Furthermore, encrypting foreign keys ensure the protection of sensitive relationships between relations involved into Inter-table constraints. However, using deterministic encryption algorithm has two issues. First, a major advantage is to enforce indistinguishability of records which allows only authorized users who know the encryption key to execute queries associating these relations. Second, a minor drawback is that it allows an adversary to infer information about repeatedly occurring values of the encrypted foreign keys, but this information

does not allow the adversary to break the unlinkability between relations. The attribute *salt* which is used as a primary key of different relations in the secure fragments protects encrypted data against frequential attacks. In addition, there is no need to secure the *salt* attribute because knowledge of the value of this attribute will not give any advantage when attacking encrypted data.

Example 4.2. Assume that we have a relational database D of a medical insurance company that contains two relations *Patient* and *Doctor* represented respectively in Table 1 and Table 2. The insurance company has defined the a set of confidentiality constraints $CC = \{C_1 = \{SSN\}, C_2 = \{Name\_pat, Illness\}, C_3 = \{Patient, Doctor\}\}$. As shown before, the first step in the fragmentation process consists in transforming Inter-table constraint (C_3). Relations *Patient* and *Doctor* are linked through the foreign key *Id_doc* in the relation *Patient*, therefore C_3 will be replaced by $C_4 = \{Id\_doc\}$ and all possible Association constraints composed of an attribute of the relation *Doctor* and an attribute of the relation *Patient* (Guarantee that the relation *Patient* will not be in the same fragment as the relation *Doctor*). In our example, attributes SSN and *Id_doc* of the relation *Patient* are involved in singleton constraints C_1 and C_4 respectively. So they will not be concerned by the fragmentation. As a result C_3 will be replaced by :

- $C_4 = \{Id\_doc\}$
- $C_5 = \{Name\_pat, Id\_doctor\}$
- $C_6 = \{Name\_pat, Name\_doc\}$
- $C_7 = \{Dob, Id\_doctor\}$
- $C_8 = \{Dob, Name\_doc\}$
- $C_9 = \{Illness, Id\_doctor\}$
- $C_{10} = \{Illness, Name\_doc\}$

Table 1. Patient relation

SSN	Name_pat	Dob	Illness	Id_doc
865746129	A. Barrett	20-08-1976	Illness_1	doc_3
591674603	C. Beat	18-01-1981	Illness_2	doc_3
880951264	N. Baines	14-09-1986	Illness_1	doc_2
357951648	S. Brandt	18-01-1981	Illness_3	doc_1

Table 2. Doctor relation

Id_doctor	Name_doc
doc_1	C. Amalia
doc_2	D. Annli
doc_3	P. Amadeus

A possible fragmentation of D that satisfies all confidentiality constraints is the set of fragments $\{F_1, F_2, F_3\}$ with: $F_1 = \{Patient(Name\_pat, Dob)\}$, $F_2 = \{Patient(Illness)\}$ and $F_3 = \{Doctor(Id\_doctor, Name\_doc)\}$. Next step is the $Secure fragmentation$ transformation (Definition 3). We assume that encryption of the protected attributes uses the deterministic encryption algorithm E with the encryption key K. The result of applying the $SecureFragmentation$ over different fragments is represented as follows.

- $F_1 : Patient(\underline{salt}, enc, Name\_pat, Dob, E_k(Id\_doc))$ with $enc = E_K(\langle SSN, Illness\rangle \bigoplus salt)$
- $F_2 : Patient(\underline{salt}, enc, Illness, E_k(Id\_doc))$ with $enc = E_K(\langle SSN, Name\_pat, Dob\rangle \bigoplus salt)$
- $F_3 : Doctor(Id\_doctor, Name\_doc)$

Note that F_3 has not been changed because there is no singleton constraints over the *Doctor* attributes. Lastly data fragments F_1, F_2 and F_3 are distributed to different Cloud storage providers.

5 Query Execution Model

Before discussing techniques for processing query over distributed fragments, we will first present the architecture of our proposed approach. It includes three principal entities : (1) a $User$ which attempts to execute queries over the original database, (2) a *Client* which rewrites user queries by splitting them to create an optimized distributed Query Execution Plan QEP; QEP is a set of sub-queries and other operations (e.g., decryption, join...), it is created by the $Query\ Transformative$ based on the $MetaData$ which contains information (relations, clear attributes, encrypted attributes, selectivity of attributes) about data distribution in different fragments. Furthermore, the *Query Executor* executes each of these sub-queries over the appropriate fragments and sends back the results to the client. (3) *Server* represented by different Cloud storage providers in which data fragments are distributed.

Query Transformation and Optimization. In our querying model, query transformation is performed by the *Query Transformative* (QT) entity on the client side. When receiving a user query, the query is analyzed syntactically and semantically so that incorrect queries are rejected as earlier as possible. Next, based on the $Metadata$ stored on the client side, the QT will attempt to find a fragment on which the user query can be executed, i.e. a fragment in which QT can find all attributes and relations involved in the user query. If such a fragment does not exist, QT will decompose the user query into queries expressed in relational algebra, find out which fragments are involved in the query, and finally transform the user query into a set of fragments queries. Using this set of fragment queries and other operations such as encryption, decryption, join and aggregation, the QT creates a QEP and sends it to the *Query Executor*. A query can have more that one QEP. Logically, each QEP may have a different execution cost. Thus, the QT should have the capability to pick out the best QEP in terms of execution cost. This capability is explained later in the Query Optimization section.

For multi-fragment query[2], QT will use local join operations as it should combine results of execution of subqueries over fragments. There are two different ways to perform local join operation : (1) Execute all sub-queries in a parallel manner, then join the result on the client side. (2) Execute sub-queries in a sequential manner to have the ability to perform *semi-joins* using the result of previous sub-queries. While (1) can be cheaper than (2) in terms of sub-query execution, it is much more costly in the join operation because in (1), sub-queries results might contain a lot of records that will not be part of the final results.

[2] i.e. a query that cannot be executed over only one fragment.

In addition to traditional query optimization methods such as selecting conditions as earlier as possible, the QT attempts to minimize the execution cost of the created QEP by applying the selection condition with the most selective attribute, i.e the selection condition which is satisfied by the smallest number of tuples. To give this ability to the QT, we assign a selectivity[3] to each attribute contained in the original database to the $Metadata$ stored in the $Client$. We apply the optimization method to the example below.

6 Preserving Data Unlinkability

Ensuring data confidentiality is achieved by preserving unlinkability between different data fragments and by encrypting all sensitive information that cannot be protected using only fragmentation. However, we have seen in the previous section that evaluation of some queries may use *semi join* in order to join data from different fragments. This will not be a concern in the case of non-colluding Cloud storage providers, but it becomes a serious concern when Cloud Storage Providers (CSP) can collude. In this section, we present our solution to overcome this privacy concern when we assume that CSP can collude to link data stored in different fragments.

To overcome this problem, the *Client* should have the ability to execute *semi join* queries and retrieve data from a fragment without the CSP (which stores the fragment) learning any information about the *semi join* condition values. To meet this requirement, we will use a Private Information Retrieval keyword-based technique. PIR keyword-based was presented in [2] to retrieve data with PIR using keywords search over many data structures such as binary trees and perfect hashing. Later, [12] investigated the use of SQL for PIR, based on the use of B+ tree. In the next part of this paper, we will explain how we can use technique presented in [12] to ensure our *semi join* queries privacy requirement.

Theorem 2. Let $\mathcal{D}$ be a multi-relation normalized database, $\mathcal{F} = \{F_1, F_2\}$ be a fragmentation of $\mathcal{D}$, and Q be a multi-fragment query that joins records from both fragments F_1 and F_2. Consider that SCPs in which the fragments F_1 and F_2 are stored can collude to link data stored in these fragments, and that Q is evaluated using *semi join* operations. Sensitive relationships between F_1 records and F_2 records remain protected if and only if the privacy of the *semi join* subqueries is guaranteed.

PIR System Design. In the *Client* of our architecture, we give to *Query Executor* the ability to communicate with different Cloud storage providers through the PIR keyword-based protocol. In the *Server*, we add on each CSP a *PIR Server* as a front-end entity to answer *Query Executor*'s PIR queries. An adversary (a Cloud storage provider administrator) who can observe *Query Executor*'s PIR-encoded queries is unable to find out the clear content of the queries. Enforcing integrity on the *PIR server* side is straightforward since we

[3] Provides an approximation of the number of tuples that satisfies a predicate.

assume that PIR servers will not attempt to wrongly answer *Query Executor*'s PIR queries.

Keyword Index Structures. The main purpose for using PIR keyword-based is to ensure the privacy of *semi join* queries. In our approach, this kind of queries is mainly executed over primary or foreign key attributes. Therefore each *PIR Server* will create a B+ Tree over each indexed attribute (Primary or foreign key). The advantage of such an index structure is that data appears only in the leaves while other nodes are used to control the search. On the leaves of B+ tree and for each key in the tree, we store the tuple corresponding to the key as the data part linked to the key. We consider the act of retrieving a node's data as a PIR operation over all nodes in the B+ tree. In all existing PIR schemes, a common assumption is that the client should know the address of the block or the item to be retrieved. To satisfy this assumption in our approach, the *PIR server* after creating the index structure, sends an index helper containing the B+ tree's root node to the client.

Semi-Join PIR Keyword-based Query. Using the PIR keyword-based query requires a setup phase in which the *Query Executor* and the *PIR server* exchange information. This setup phase is divided into two steps:

1. The *Query Executor* sends the Relation schema name and the attribute name over which the *semi join* query is to be performed to the corresponding *PIR server*.
2. When receiving the Relation schema name and the attribute name, the *PIR server* selects the corresponding B+ tree and sends its root node to the *Query Executor*.

After receiving the root node sent by the *PIR server*, the *Query Executor* will compare the list of keys contained in the root node with values used in the condition of the *Semi join* query in order to find out the indexes of the next nodes to be retrieved. The *Query Executor* will subsequently perform PIR queries over chosen indexes to retrieve corresponding nodes. Once all items have been retrieved, The *Query Executor* combines them to build the result of the original *Semi join* query. Refer to Appendix D for a description of the PIR keyword-based protocol algorithms used in the *Client* and the *Server* parts.

7 Conclusion

Existing approaches based on fragmentation and encryption have focused on single-relation schema database which is a strong and rarely encountered assumption in real environment. In this paper, we have presented a new approach based on fragmentation, encryption and query privacy techniques which ensures confidentiality of outsourced multi-relation databases. We have presented also a way to optimize and execute queries over distributed data fragments.

Our future work will include the implementation of our approach. It will also include enhanced query optimization and execution techniques to overcome some limitations of our approach, such as processing nested queries.

References

1. Bellare, M., Fischlin, M., Ristenpart, T.: Deterministic encryption: Definitional equivalences and constructions without random oracles (2008)
2. Benny Chor, N.G., Naor, M.: Private information retrieval by keywords. Cryptology ePrint Archive, Report 1998/003 (1998)
3. Biskup, J., Preuß, M., Wiese, L.: On the Inference-Proofness of Database Fragmentation Satisfying Confidentiality Constraints. In: Lai, X., Zhou, J., Li, H. (eds.) ISC 2011. LNCS, vol. 7001, pp. 246–261. Springer, Heidelberg (2011)
4. Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Fragmentation and Encryption to Enforce Privacy in Data Storage. In: Biskup, J., López, J. (eds.) ESORICS 2007. LNCS, vol. 4734, pp. 171–186. Springer, Heidelberg (2007)
5. Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Fragmentation design for efficient query execution over sensitive distributed databases. In: ICDCS, pp. 32–39. IEEE Computer Society (2009)
6. Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Livraga, G., Samarati, P.: Enforcing Confidentiality and Data Visibility Constraints: An OBDD Approach. In: Li, Y. (ed.) DBSec. LNCS, vol. 6818, pp. 44–59. Springer, Heidelberg (2011)
7. Hacigümüs, H., Iyer, B.R., Li, C., Mehrotra, S.: Executing sql over encrypted data in the database-service-provider model. In: SIGMOD Conference, pp. 216–227. ACM (2002)
8. Hacigümüs, H., Mehrotra, S., Iyer, B.R.: Providing database as a service. In: ICDE, pp. 29–38. IEEE Computer Society (2002)
9. Hudic, A., Islam, S., Kieseberg, P., Weippl, E.R.: Data confidentiality using fragmentation in cloud computing. International Journal of Communication Networks and Distributed Systems, IJCNDS (2012)
10. Ponemon Institute. Fourth annual us cost of data breach study (January 2009)
11. Ceri, S., Wiederhold, G., Navathe, S.B., Dou, J.: Vertical partitioning of algorithms for database design. ACM Trans. Database Syst. 9(4), 680–710. 98, 99, 102, 109, 125 (1984)
12. Olumofin, F., Goldberg, I.: Privacy-Preserving Queries over Relational Databases. In: Atallah, M.J., Hopper, N.J. (eds.) PETS 2010. LNCS, vol. 6205, pp. 75–92. Springer, Heidelberg (2010)

A Secure Fragmentation Algorithm

Algorithm 1. Secure fragmentation

Require:
$\mathcal{D} = \{R_1, R_2, \cdots, R_n\}$ /* Normalized relational database */
$\mathcal{C} = \{C_1, C_2, \cdots, C_m\}$ /* Confidentiality constraints */
Ensure:
$\mathcal{F}^s = \{F_1^s, F_2^s, \cdots, F_p^s\}$ /*The set of secure fragments*/
Main
$\mathcal{C}_f = \{C_i \in \mathcal{C} : |C_i| > 1\}$ /* The list of association constraints */

$\mathcal{A}_{fkey} = \{a \in C_i, C_i \in \mathcal{C} : |C_i| = 1 \; and \; \textbf{isForeignKey}(a) = True\}$
/* $\mathcal{A}_{fkey}$: The set of foreign keys to be encrypted*/
$\mathcal{F} := \textbf{Fragment}(\mathcal{D}, \mathcal{C}_f)$
for all $F_i = \{a_{i_1}, a_{i_2}, \cdots, a_{i_n}\}$ in $\mathcal{F}$ **do**
  $\mathcal{R}_f = \textbf{classifyAttributes}(F_i)$ /* Classify attributes*/
  for all R_{f_i} in $\mathcal{R}_f$ **do**
    for all r in R_{f_i} **do**
      $r^s[salt] := \textbf{GenerateSalt}(R_{f_i}, r)$ /* r : record */
      $r^s[enc] := \mathcal{E}_k(t[a_{j_1}, \cdots, a_{j_q}] \oplus r^s[salt])$ /* $a_{j_1}, \cdots, a_{j_q} = R_i - R_{f_i}$ */
      for all a in R_{f_i} **do**
        $r^s[a] := r[a]$ /* a : attribute */
      end for
      for all a in A_{fkey} **do**
        if $a \in R_i$ **then**
          $r^s[a] := \mathcal{E}_k(r[a])$ /* a : the foreign key of the relation R_i */
        end if
      end for
      InsertRecord(r^s, R^s)
    end for
    AddRelationToFragment(R^s, F^s)
  end for
end for

B Proof of Theorem 1

Proof. According to Item (ii) of Definition 2, the Inter-table constraint will be replaced by all possible associations constraint composed from an attribute of relation R_1 and another from relation R_2. Due to the fact that an association constraint between two attributes means that the relationship between these attributes will be protected using fragmentation (each attribute will be stored in different fragments), Item (ii) guarantees that relations R_1 and R_2 will be stored in different fragments which hold condition (2).

Item (i) of Definition 2 creates a singleton constraint over the foreign key b_m of the relation R_2. Thus b_m will be considered as a sensitive attribute and will be protected using encryption, which means that the foreign key b_m will not appear in clear in any fragment. As a result, if an adversary succeeds in having access to the fragments in which R_1 and R_2 have been stored, she is unable to link data stored in these relations.

C Proof of Theorem 2

Proof. To prove the Theorem 2, we will use the following two sketches. The fist sketch proves that without ensuring *semi join* sub-queries privacy, collaborative CSPs can, in some cases, break data unlinkability, while the second sketch proves that, under a collaborative Cloud storage service providers model,

protecting data unlinkability can only be guaranteed with the protection of the privacy of the *semi join* sub-queries.

SKETCH Without Using the PIR Keyword-based Protocol. Suppose that the *Cient* wants to execute a query which joins records from two fragments F_1 and F_2. Let us consider that the sub-query Q_1 executed over the fragment F_1 has returned n tuples. And the semi-join query Q_2 executed over F_2 has returned m tuples. Therefore, if CSPs that store F_1 and F_2 collude together to link tuples from Q_1 and Q_2 results, the probability to guess correctly the relationship between tuples is:

$$PROB[Result(Q_1) \leftrightarrow Result(Q_2)] = \frac{1}{m \times n}$$

Clearly, if m and n are small, CSPs will have a great chance to break data unlinkability.

SKETCH Using the PIR Keyword-based Protocol. Let us consider that the *Client* attempts to perform a query which joins records from two fragments F_1 and F_2. According to our defined PIR keyword-based protocol, the *Client* will execute Q_1 over the fragment F_1 without using the keyword-based protocol. Next, the *Client* will send the table name T and the attribute name a on which the semi-join will be performed, the *Server* replies with the root node of the corresponding B+ tree. It is clear from the previous step that the CSP which stores F_2 can only know the attribute name and the table name on which the semi-join will be performed. After receiving the root node, the *Client* will use the PIR protocol to retrieve internal corresponding nodes until the leaves of the B+ tree are reached. The PIR protocol will ensure that the server will not know which nodes were retrieved by the *Client*. Moreover, all tuples are stored in the leaf level of the B+ tree. Therefore, in order to retrieve each record, the *Client* shall execute the same number of PIR queries. Rightfully, the only revealed information when using the PIR keyword-based protocol is the table name and the attribute name on which the semi-join has been performed. Therefore, if CSPs storing F_1 and F_2 collude together to break data unlinkability, they will be able only to infer that the relation T_1 in F_1 over which Q_1 has been executed is linked to the relation T through the attribute a. Due to the fact that the foreign key in T_1 referencing the attribute a in T is encrypted, linking records is not possible.

D SemiJoin PIR Keywordbased Query Algorithms

Algorithm 2. SemiJoin PIR keywordbased query (server)

Require:
$BPT = \{B_1, \ldots, B_n\}$ /* B-Plus Tree over indexed attributes*/
loop
$Request \leftarrow handle\_client\_request()$
if $Request$ is PQR **then**
/* PQR : Pre-Query Request */
$(TabName,\ AttriName) \leftarrow Request$
$B \leftarrow GetAssociatedBPT(TabName,\ AttriName)$
$Root_B \leftarrow GetRootNode(B)$
$ReplyToClient(Root_B)$
else {$Request$ is PIRQ}
/* PIRQ : PIR Query */
$result \leftarrow compute(Request)$
$ReplyToClient(result)$
end if
end loop

Algorithm 3. SemiJoin PIR keywordbased query (client)

Require:
$tabName, attrName$ /* Table and Attribute where the semi-join will be performed*/
$value$ /* Semi-join condition value*/
$Node \leftarrow send\_PQR\_request(tabName, attrName)$
repeat
for all $elem$ in $Node$ **do**
$findLink \leftarrow$ **false**
if $Key(elem) < value$ **then**
$Node \leftarrow PIR\_Query(IndexOfLeftChild(elem))$
$findLink \leftarrow$ **true**
break /* terminates the for loop*/
end if
end for
if $findLink =$**false then**
$Node \leftarrow PIR\_Query(IndexOfRightChild(elem))$
end if
until $Node$ is leaf_node
for all $elem$ in $Node$ **do**
if $Key(elem) = value$ **then**
return $Data(elem)$
end if
end for

Enhancing Privacy Protection in Distributed Environments through Identification and Authentication-Based Secure Data-Level Access Control

Nisreen Alam Aldeen and Gerald Quirchmayr

University of Vienna
Faculty of Computer Science
Research Group Multimedia Information Systems
Währinger Straße 29, 1090 Wien, Austria
a0948830@unet.univie.ac.at, gerald.quirchmayr@univie.ac.at

Abstract. System-level access control methodologies depending on Perimeter Protection proofed their efficiency in the past, but the appearance of many new significant developments in digital communications highlighted the limitations of this approach. Increased concerns about the compatibility of system-level access control mechanism with new distributed and ubiquitous environments are turning aspirations towards de-perimeterisation protection and data level access control as solutions. This research does therefore try to make a contribution to privacy protection based on already advanced data-level access control work, such as the SPIDER project. The solution developed in this research suggests an X.509 certification extension to fit the data-level access control requirements, and proposes a new design for application structure in order to improve the identification and authentication-based secure data-level access control process.

Keywords: Privacy Protection, De-perimeterisation (De-P), Data-level Access Control, Self-Protecting Data, X.509 certification.

1 Introduction

Perimeter protection represented by system-level access control was the earliest framework to provide and support privacy protection and handle control. Despite the great implementation features of system-level access control there are various limitations helped to raise awareness related to the problems of perimeter protection, and to promote De-perimeterisation (De-p) protection [6]. The main purpose of De-p protection is to attain continuous and modifiable access control to the information shared beyond the system's boundaries [3]. Digital rights management (DRM) with its security processes can be considered as a non- perimeter security model, where the control can be applied to the resource outside the system's boundaries. DRM access control still is static, not modifiable, and a fine-grained classification scheme is missing [1]. The JISC funded SPIDER project (Self-Protecting Information for De-Perimeterised

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 441–446, 2013.
IFIP International Federation for Information Processing 2013

Electronic Relationships) [2] is one of the currently very promising solutions based on the De-p principle. It aims to provide a set of tools including classification scheme, persistent and modifiable access control and enforcement mechanisms for information shared in collaborative distributed work environments. However, there are specific limitations, related to identification and authentication issues, which restrict its efficiency and put a stop to its reliability.

The aim of this paper is to attempt to bridge the gap between the SPIDER project and DRM techniques, using modifiable digital certification construction. More specifically, in this paper we suggest an extended X.509 certificate [10] to be compatible with modifiable and continuous data level access control in distributed environments, in addition to a modified SPIDER application that allows shifting the emphasis from system as controller to user-definable policies. Starting with an overview of existing technologies the paper identifies the most interesting gaps and then explains the own research approach, its expected benefits and limitations, it concludes with a reflection of the current state of the work and outlook to future work.

2 Existing Access Control Technologies

System- level access control is the traditional strategy of existing access control systems. The central theme of this strategy is perimeter protection which means the application of control to information access requests within the system's boundaries using previously assigned access rights. System-level access control provides a sufficient access control and authentication infrastructure for entities. However, it has various limitations [1] such as a lack of information classification schemes, and a Lack of continuous and modifiable control behind the secure network boundaries.

Furthermore, the emergence of many new variables and environments in the context of digital communication (e.g., global corporate, cloud computing, and online business) highlighted the problem of system-level access control, and showed the inability of the perimeter protection to be compatible with the new environments [4]. Growing problems of level access control associated with disappearing boundaries between networks have raised an important question: What is the alternative? And how can we extend the control on our data beyond the system's boundaries?

2.1 De-perimeterisation Protection as a Key

The first description of the De-p protection was in the Jericho forum [5] [7]. It suggests shifting from protecting data from the outside (system and applications) to protecting data from within. Mainly it aims to apply continuous control to data outside the system's boundaries using data-level access control (i.e., Information-centric security) and including some access control policies into the information itself [3].

2.2 Some Existing Approaches for De-perimeterisation Protection

Many methods have been presented in order to solve the problems of perimeter protection, and they can be viewed as partial solutions towards de-p achievement. Partly

we can look at digital rights management DRM as non- perimeter security model as it provides a continuous control beyond the system's boundaries. However DRM access control is static, not modifiable and doesn't provide a fine grained classification scheme. In 2001 and 2002, two selective access approaches for information sharing in distributed environments via the Web have been presented [8] [9]. These two approaches aim to protect information using XML's Properties to allow the fragmentation of content, and to apply different controls to the same resources. The major drawback of these approaches is the lack of persistent and modifiable access controls. The SPIDER project [1] [2] being led by Cardiff University combines the advantages of previous methods. It provides a practical way to modify the existing models of information classification based on access control requirements, with a live URL link which allows the possibility to enforce and modify the access control policy beyond the system boundaries. The SPIDER Project represents a valuable approach towards solving the Perimeter protection problems, certainly, the problem of non-continuous and non-modifiable control. However, a few aspects of SPIDER are still under development, and there are considerable gaps effect on SPIDER performance and efficacy.

2.3 Building on Existing Work

With the SPIDER project [2] already covering the core functionalities, we can focus on closing the following interesting gaps:

- Trusted computing: With the application of SPIDER, the entire resource is encrypted with the same key for all security levels; then the whole resource in memory is open to attack in case of key disclosure. In addition, a trusted computing module to ensure confidentiality during decryption encryption is required.
- Identity management is still under development, there is a trend to use a digital certificate (not yet defined).

Our previous description of the SPIDER limitations highlights the need for confidentiality, authentication and identity management mechanisms.

3 Towards a Solution Model

Following on from this, a set of requirements based on the drawbacks of the previously mentioned methods will be derived, then used to implement a solution that considers these requirements using a modified application of SPIDER [1], in order to provide a solution that enhances the access control process in widely distributed and expanding collaborative working environments, and to improve the information exchange privacy by providing X.509 extension to allow fine grained access controls.

Taking advantage of the X.509 standard certificate [10] which can be viewed as one of the DRM processes, and taking into account the requirements of data-level access control architecture such as SPIDER project [2], using X.509 certificate to enhance the confidentiality and authentication is not sufficient. Despite the high level of confidentiality achieved by X.509 there are significant drawbacks affecting X.509 adoption in distributed structures:

- The user is not a controller; everything is done by a trusted third party.
- Doesn't provide a classification scheme for access control
- The access control depends on previously assigned rights; thus, it is not modifiable.

On the other hand, SPIDER doesn't provide authorization credentials or authentication mechanisms. In addition, we still have to deal with the crisis of using one key to encrypt the entire resource. Therefore, in order to enhance access control process reusing the De-p structure of the SPIDER, we suggest a new vision as follows:

- Combining the advantages of digital certification and the SPIDER project.
- Mixing a centralized strict hierarchical structure with a stand-alone structure.
- Using various keys to encrypt the resource for different levels of security.
- Suggesting a modified X.509 digital certificate ISAC (Identification and Secure Access Control Certificate) for the purpose of identification, authentication and secure data-level access control.

3.1 Suggested Digital Certificate

The suggested certificate is generated centrally by a trust authority and contains the original structure of the X.509 [10], in addition to a new annex contains compatible fields for secure data-level access control process (as figure.1 shows). At the time of issuing an extended X.509 certificate, the center applies its electronic signature on the main part of the X.509. The special annex should later be signed by the resource's owner if he wishes to grant access privileges without referring to the authority center.

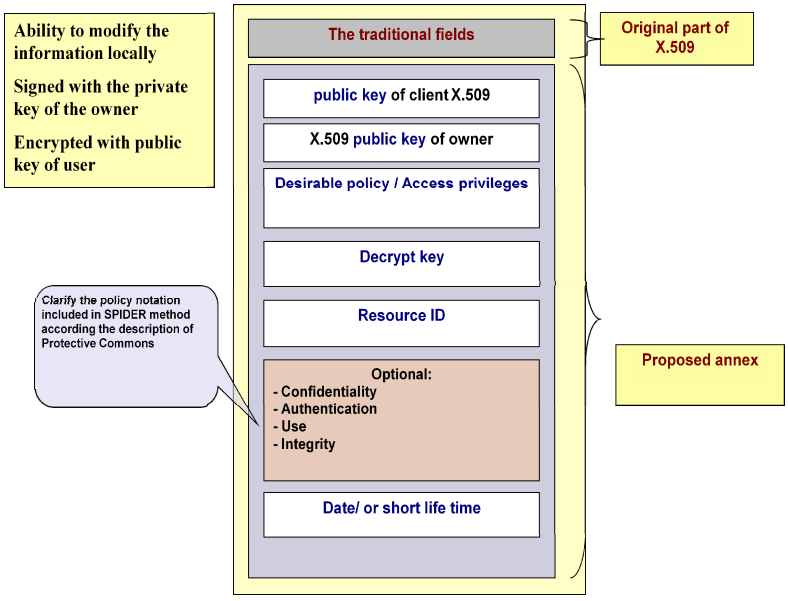


Fig. 1. The essential fields of the suggested annex

3.2 Basic Structure of the Process

The underlying idea for a process is illustrated in figure.2, when the user wishes to access the resource; he will send an access request including his certificate. The owner side will check the identity and connect the database to extract the policies and keys which match this identity; then it will fill the fields of the annex with the required information, sign it with the private key of the owner and encrypt it using the public key of the user before sending it back to the user. The SPIDER application in the user side will verify the access control annex, parse the resources for the classifications labels that match the security label returned for the user and generate a dynamic subset of the original resource in unencrypted form for the user to access.

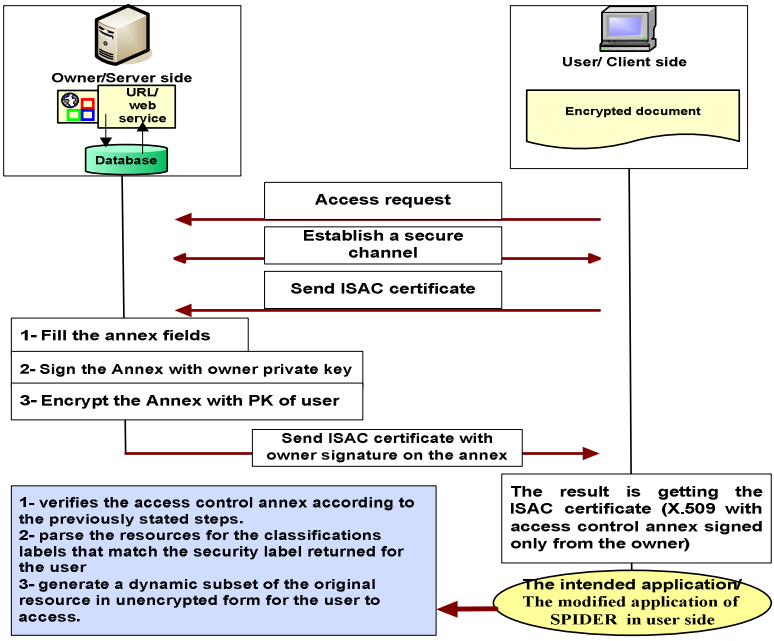


Fig. 2. Basic structure of the process

3.3 Advantages and Limitations of the Suggested Solution

The suggested solution exploits the high level of security in X.509, gaining the advantages of the hierarchical structure in confidentiality and authentication. Including access rights in the extended certificate enables the use of classification methods and partial access rights, these rights are continuous and modifiable by drilling down to the level of data using SPIDER project. The suggested X.509 extension allows this certificate to be used in distributed environments and paves the way for building a standards-based for de-p approach in the future. Furthermore, the use of different keys for encrypting the resource protects the encrypted resource existing in memory in case of key disclosure. The suggested solution has some well-known limitations, including limiting data-level access control operations to X.509 holders, and using various keys

to encrypt the resource, which takes us back to the key management problem, for which we should measure the efficiency of our suggestion in reality.

4 Conclusions and Future Work

The inability of perimeter protection and system-level access control mechanism to match new significant developments in digital communications made it a must to come up with a new flexible and secure access control method. In this paper we have provided a brief description of the SPIDER project as the currently most advanced and practical method based on De-perimeterisation protection and data-level access control. We have investigated the X.509 standard certification and highlighted its disadvantages focusing on the difficulties of applying this standard to distributed environments to be completed in our current research. Later in the paper we have introduced our suggestion aiming at two improvements:

- Strengthening the authentication and identification mechanism of SPIDER project
- Making the X.509 standard certificate applicable and more efficient in a distributed environment by adding a data-level access control annex.

A deeper security analysis of this extended certificate is planned and experiments will be conducted to test the viability of this approach in realistic settings.

References

1. Burnap, P., Hilton, J.: Self Protecting Data for De-perimeterised Information Sharing. In: Third International Conference on Digital Society. Cardiff School of Computer Science. Cardiff University (2009)
2. Burnap, P., Hilton, J., Tawileh, A.: Self-Protecting Information for De-perimeterised Electronic Relationships (SPIDER), Final Report (July 30, 2009)
3. van Clee, A., Wieringa, R.: De-perimeterisation as a cycle: tearing down and rebuilding security perimeters. (December 5, 2008)
4. Olovsson, T.: CTO, The Road to Jericho and the myth of Fortress Applications, AppGate Network Security, Appgate (2007)
5. Olovsson, T.: Surviving in a hostile world, The myth of fortress applications, CTO, Appgate, Jericho Forum, Professor at Goteborg University (2007)
6. Palmer, G.: De-Perimeterisation: Benefits and limitations, Network and Security, Siebel Systems Ltd., Siebel Centre, The Glanty, Egham, Surrey TW20 9DW, United Kingdom (2005)
7. Simmonds, P.: Architectures for a Jericho Environment, Global Information Security Director, ICI Information Security (2004)
8. Damiana, E., De Cabitani Di Vimercati, S., Paraboschi, S., Samarati, P.: A Fine-Grained Access Control System for XML Documents. ACM Transactions on Information and System Security 5(2), 169–202 (2002)
9. Bertino, E., Castano, S.: On Specifying Security Policies for Web Documents with an XML-based Language (2001)
10. ITU-T Recommendation X.509, ISO/IEC 9594-8, Information Technology – Open Systems Interconnection- The Directory: Authentication Framework (1997)

Toward Secure Clustered Multi-Party Computation: A Privacy-Preserving Clustering Protocol

Sedigheh Abbasi, Stelvio Cimato, and Ernesto Damiani

DI - Università degli Studi di Milano, 26013 Crema, Italy
firstname.lastname@unimi.it

Abstract. Despite a large amount of research work has been done and a large number of results produced, the deployment of Secure Multi-party Computation (SMC) protocols for solving practical problems in real world scenarios is still an issue. This is mainly due to the complexity of the SMC-based solutions and to the needed assumptions that are not easy to fit to the considered problem. In this paper we propose an innovative approach for the deployment of SMC, providing a tradeoff between efficiency and privacy. In the Secure Clustered Multi-Party Computation (SCMC) approach, a function is more efficiently computed through reducing the number of participants to the SMC protocol by clustering, such that a reasonable privacy leakage inside the cluster is allowed. Toward this direction, this paper verifies the impact and the feasibility of applying different clustering techniques over the participants to a SMC protocol and proposes an effective specifically-tailored clustering protocol.

Keywords: Secure Multi-Party Computation, Privacy-Preserving Clustering, Privacy and Efficiency Tradeoff.

1 Introduction

Secure computation was introduced by A. Yao [6] in 1982 by presenting the solution to the millionaires' problem as a two-party computation protocol and successively generalized to include more parties [4]. In general, secure multi-party computation (SMC) refers to protocols where multiple participants want to jointly compute the value of a public function over their individually-held secret bits of information without revealing the secrets to the other participants. In such a setting, the security of the protocol is said to be preserved if no participant can learn more from the description of the public function and the result of the global calculation than what he could learn from his own input (under particular conditions and depending on the model used). In other words, in a SMC protocol the only information that should be revealed is what can be reasonably deduced by knowing the actual result of the function. As a fundamental theorem, it has been proved that any multi-party functionality can be securely computed [1, 3, 4, 6].

However, in many cases, these significant achievements are not so efficiently scalable and adaptable to provide a solution to real-world problems. Indeed, few practical applications of SMC protocols are reported in literature, such as the first large-scale

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 447–452, 2013.
IFIP International Federation for Information Processing 2013

practical experiment, where SMC is used to implement a secure sugar-beet auction in Denmark [2] and the usage of SMC to supply chain management, where the shared optimization problem is solved by maintaining the privacy of the user inputs [5]. In both cases, the complexity of the solutions is huge and the completion time of the protocol is directly proportional to the number of participants, limiting the applicability of the proposed solutions.

In order to achieve a more scalable solution, in this paper, we propose to execute SMC protocols over a significantly-reduced number of participants by appropriately clustering the participants. The paid price is in terms of privacy, since intra-cluster computations will rely on a relaxed (or none) privacy model while the SMC protocol will be executed over clusters' representatives among which privacy assumptions will be preserved. This approach in which a SMC protocol is executed over the participants' clusters rather than individual participants is called Secure Clustered Multiparty Computation (SCMC) and is depicted in Fig. 1.

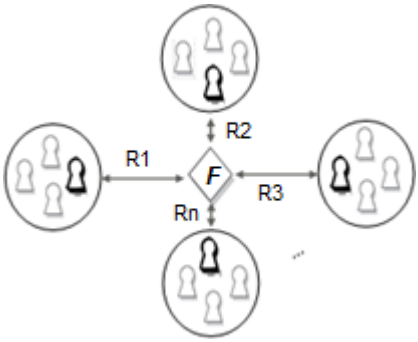


Fig. 1. An abstract view of a SCMC protocol. Bold symbols show clusters' representatives and R1 to Rn denote the results of intra-cluster computations.

In short, the overall execution of a SCMC protocol can be divided in two phases: clustering phase and computation phase. During the clustering phase participants are clustered in such a way that a more efficient intra-cluster computation is possible, by applying a relaxed privacy model. Throughout this paper, the relaxed privacy model we adopt is a trust-based model, where privacy leakages from a confiding participant to a trustee are allowed. Therefore, in its simplest representation, it can be assumed that all the clustering decisions are made based on a square trust matrix (T) in which the element $T_{i,j}$ denotes the amount of trust of participant P_i to the participant P_j. On the other hand, the clustering algorithm itself should be privacy-preserving; because, it uses pair-wise trust values as inputs that are private to the participants. Then, the computation phase starts with intra-cluster computation to evaluate partial results of the clusters and ends with a SMC protocol execution run over the clusters' representatives as participants, holding clusters' partial results, as private inputs.

The purpose of this paper is to address the issues related to the participants' clustering in a SCMC protocol. Clearly, these issues are independent of the specific SMC protocol that is run in the last phase of the protocol and therefore all discussions about

different SMC options, their underlying assumptions and definitions remain unchanged and are outside the scope of our study.

2 A Privacy-Preserving Clustering Protocol in SCMC

In order to achieve an efficient clustering, we propose a general privacy-preserving clustering protocol that is particularly designed to work with the asymmetric trust matrix as input and do not impose any limitation over the representation of trust values. In other words, each participant is independently free to evaluate her trust to the other participants. This protocol, which is similar to a voting system, securely defines the clusters in terms of balanced hierarchies. During the execution of the protocol, the participants are prioritized and the owners of the highest priorities are introduced to the others as candidates for the current election. On the other hand, the protocol utilizes a Trusted Third-Party (TTP) to improve efficiency and prevent possible privacy leakages.

2.1 TTP-Based Prioritized Voting Clustering Protocol

As mentioned above, the Prioritized Voting Clustering (PVC) protocol is in fact a prioritized voting system in which an election is run during each iteration for one or more participants with the highest priorities as the current candidates. A candidate is elected by a participant only if the participant agrees to share her private input with the candidate in the computation phase of the SCMC protocol. In this situation, it is said that the participant has chosen to be a follower of the candidate. Such a successful election leads to the formation of a parent-child relation in the clustering hierarchy in which the parent represents the leader. The candidates are in fact chosen in such an order that favors the creation of the minimum number of balanced clusters, taking into account two factors affecting the participants' priorities: the probability of becoming a follower (a child) and the current level in the clustering hierarchy (all the participants start the protocol at level 0). The first probability can be simply estimated by considering the number of trustees of each participant (the less the number of trustees is, the more probably the participant is a parent). Also, in order to prevent skew clustering, the participants in the lower levels of the clustering hierarchy (closer to the root) are also given higher priorities. Therefore, the priority of a participant is defined as the sum of her level in the clustering hierarchy and her number of trustees: the less the result of this sum is, the higher priority the participant has. Note that, for the sake of simplicity, here a uniform binary trust matrix has been assumed (even if our protocol does not put any limitation on the trust representations used by different participants). On the other hand, since a cluster hierarchy is not allowed to have any cycle, a participant cannot vote for one of its ancestors in the clustering hierarchy. For this reason, all the participants are required to keep track of their followers.

This algorithm preserves the privacy of the private trust vectors in multiple ways: Firstly, by introducing a TTP, secondly by using one of the participants (namely P0) as the inductor and thirdly by utilizing a public-key cryptosystem that is used to hide the status of each participant from the other participants as well as from the TTP. In fact, the algorithm benefits from the TTP not only to prevent the disclosure of the trust values among the participants, but also to create a central coordinator that helps

to improve the efficiency of the clustering by reducing the overall number of needed elections (since more than one candidate in an iteration is considered when possible). On the other hand, in order to preserve the participants' privacy against the TTP itself, the inductor is used: any information from the participants is sent to the TTP after being collected by the inductor, while the results from the TTP are directly broadcast to all the participants. As a result, due to this obfuscation, the participants are kept anonymous to the TTP.

Algorithm 1 illustrates the initialization and the execution phases of the protocol. Note that, this algorithm requires that all the participants are identified by a unique ID that is known to the other participants and kept unknown to the TTP. In this way, the TTP can access the public keys, but it cannot recognize whom a particular key belongs to.

Algorithm 1. TTP-Based PVC Algorithm

Initialization Phase:

1) Each participant marks itself as $Not\ Connected$.
2) All the participants send a tuple $< Enc_{PK}(ID), Enc_{TK}(NOT), PK >$ to the inductor, where PK is the public key of the participant, $Enc_{PK}(ID)$ denotes the encryption of its ID under its public key, and $Enc_{TK}(NOT)$ is the encryption of the number of its trustees under the public key of the TTP.
3) The inductor prepares a randomly permuted list collecting the tuples from all the participants, including him and sends the list to the TTP.
4) The TTP initializes n disjoint sets representing the initial clusters, one for each of the received tuples, using $Enc_{PK}(ID)$ as index. For each tuple he initializes the following variables:

$$ID = Enc_{PK}(ID), PublicKey = PK, Level = 0, Priority = NOT + Level, Elected = 0$$

Execution Phase:

1) The TTP finds the participants with the highest priorities whose $Elected = 0$ (as the candidates of the current election) and performs the following steps:
   1-1) For all the candidates sets $Elected$ to 1.
   1-2) Runs an election by broadcasting the list of candidates' $Enc_{PK}(ID)$s.
2) All the participants do as the following:
   2-1) If he is marked as $Not\ Connected$ and he trusts any of the candidates, he selects one of the candidates and sets her vote as $V = Enc_{PK}(ID')$ where ID' is the identifier of the selected candidate; otherwise the vote is set to $NULL$.
   2-2) The tuple $<Enc_{PK}(ID), Enc_{TK}(V) >$ containing the encrypted vote is sent to the inductor.
3) The inductor collects all the tuples from the participants as well as her own, perform a random permutation and sends the list to the TTP.
4) The TTP decrypts the second element of each pair in the list and extracts the vote (V) corresponding to a particular $Enc_{PK}(ID)$. Then the following operations are done:
   4-1) If $V = NULL$, the result R is set to $NULL$, otherwise $V = Enc_{PK}(ID')$. If $Enc_{PK}(ID')$ and $Enc_{PK}(ID)$ belong to different clusters R is set to $Enc_{PK}(ID')$, the two clusters are merged and $Level$ and $Priority$ variables related to $Enc_{PK}(ID)$ are updated.
   4-2) For each of the received pairs an ordered-pair in the form of $(Enc_{PK}(ID), Enc_{PK}(R))$ is created in which $Enc_{PK}(R)$ is the result encrypted under the corresponding $PublicKey$. These ordered-pairs form the $Results$ list that is broadcast to all the participants.
5) Each participant who has voted in the current election decrypts the corresponding $Enc_{PK}(R)$ using her private key and if R is not $NULL$, he can retrieve the ID that identifies the her leader (her parent in the hierarchy) and marks himself as $Connected$; otherwise the message is ignored by the participant.

This protocol completes when for all the participants $Elected = 1$. In this situation, the TTP broadcasts a $Terminate$ message to all the participants. Also, the TTP may decide to terminate the protocol sooner if it recognizes that a sufficient number of clusters have been achieved or the clustering will not be changed anymore.

2.2 Algorithm Analysis

From the viewpoint of efficiency, in the worst case the protocol is completed after n iterations. This situation occurs when in each iteration only one participant with the highest priority exists. On the other hand, in the best case, the protocol needs only one iteration, that is when the trust matrix leads to the same initial priority for all the participants and, moreover, there is no conflict among the selected candidates. During each iteration $O(n)$ number of $O(nk)$ bit messages are transferred, where k is the security parameter of the underlying field or ring and n denotes the number of participants. Therefore, the overall number of bits communicated during the protocol is $O(n^3k)$ and $O(n^2k)$ in the worst and the best case, respectively. From the computation point of view, a constant number of encryptions and/or decryptions are required in each iteration.

Considering privacy, the security of this algorithm depends on the hardness of the underlying cryptosystem. In other words the protocol itself does not disclose any information about the trust values neither to the other participants, nor to the TTP. The only information revealed from the participants to the TTP is the number of trusted participants each participant has, without revealing which participant particular information belongs to (this information can be utilized by the TTP to recognize when no more connection is possible among the participants and terminate the protocol in fewer rounds).

2.3 Computation Phase

As mentioned earlier, the computation phase of a SCMC protocol can be divided in two stages: *intra-cluster* and *inter-cluster* computation. The first stage is in fact a bottom-up computation in which the participants' private inputs flow up the clustering hierarchies toward their direct parents (their selected leaders). For a parent, then, there are two different strategies: *aggregation* and *shuffling*. In the former, the parent is responsible for computing the function over the inputs collected from its children as well as its input (aggregation) and sending the result to its direct parent if any. In other words, in this computation model partial results of the function are gradually computed and propagated in the clustering hierarchy, and are finally aggregated in the clusters' roots. In the shuffling model a non-root parent does not participate in the actual computation but he is simply responsible for shuffling and forwarding a list containing the inputs as well as their children's. In this way, no non-root participant is aware of the function that is supposed to be computed in the SMC phase. The main purpose of aggregation and shuffling is to prevent *unwanted* inputs' disclosure especially in the case of non-transitive trust relations.

In any case, the intra-cluster computation is over when all the roots receive the partial results or lists of inputs from all their ancestors. Then, the roots do a local computation to evaluate the partial results of the function over the inputs included in their clusters and take part to a SMC protocol as the representatives of their followers.

3 Conclusions

In this paper we analyzed the problem of trust-based privacy-preserving clustering of participants in a SCMC protocol. After studying some of the recently-proposed solutions, we showed that none of the existing clustering protocols are fully consistent with the requirements of SCMC. In future works we plan to improve our proposed clustering protocol by eliminating the TTP and replacing it with a secure distributed protocol. In this way, it would be applicable in other contexts when access to a TTP is not possible.

References

1. Ben-Or, M., Goldwasser, S., Wigderson, A.: Completeness Theorems for Non-Cryptographic Fault-Tolerant Distributed Computation (Extended Abstract). In: The Proceedings of the 20th Annual ACM Symposium on Theory of Computing (STOC), pp. 1–10 (1988)
2. Bogetoft, P., Christensen, D.L., Damgård, I., Geisler, M., Jakobsen, T., Krøigaard, M., Nielsen, J.D., Nielsen, J.B., Nielsen, K., Pagter, J., Schwartzbach, M., Toft, T.: Secure Multiparty Computation Goes Live. In: Dingledine, R., Golle, P. (eds.) FC 2009. LNCS, vol. 5628, pp. 325–343. Springer, Heidelberg (2009)
3. Chaum, D., Crépeau, C., Damgard, I.: Multiparty Unconditionally Secure Protocols. In: The Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing (STOC), pp. 11–19 (1988)
4. Goldreich, O., Micali, S., Wigderson, A.: How to Play any Mental Game or a Completeness Theorem for Protocols with Honest Majority. In: The Proceedings of the 19th Annual ACM Symposium on Theory of Computing (STOC), pp. 218–229 (1987)
5. Kerschbaum, F., Schropfer, A., Zilli, A., Pibernik, R., Catrina, O., de Hoogh, S., Schoenmakers, B., Cimato, S., Damiani, E.: Secure Collaborative Supply-Chain Management. IEEE Computer 9, 38–43 (2011)
6. Yao, A.C.: Protocols for Secure Computations (Extended Abstract). In: The Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (FOCS), pp. 160–164 (1982)

A Real-Time Privacy Amplification Scheme in Quantum Key Distribution

Bo Liu[1], Bo Liu[1], Baokang Zhao[1,*], Dingjie Zou[1], Chunqing Wu[1], Wanrong Yu[1], and Ilsun You[2]

[1] School of Computer Science National University of Defense Technology Changsha, Hunan, China
{liub0yayu@gmail.com}, {boliu,bkzhao,chunqingwu}@nudt.edu.cn
[2] School of Information Science Korean Bible University Seoul, Korea
isyou@bible.ac.kr

Abstract. QKD (Quantum Key Distribution) technology, based on the laws of physics, can create an unconditionally secure key between communication parties. In recent years, researchers draw more and more attention to the QKD technology. Privac amplification is a very significant procedure in QKD system. In this paper, we propose the real-time privacy amplification (RTPA) scheme which converts the weak secret string W to a uniform key that is fully secret from Eve. We implement RTPA scheme based on CLIP (Cvqkd Ldpc experImental Platform) which is connected to the real quantum communication systems. Experimental results show that, our proposed RTPA scheme is very efficient when the bit error rate of quantum channel is lower than 0.06.

Keywords: QKD, privacy amplification, security.

1 Introduction

Quantum Key Distribution (QKD) [1, 2] is technology for solving the key distribution problem. QKD system, based on the laws of physics, rather than the computational complexity of the mathematical problems assumed by current cryptography methods, can create an unconditionally secure key between communication parties. These keys are generated over unsecured channels, where may exist an active computationally unbounded adversary Eve.

After the procedure of information reconciliation [3], Alice and Bob have own almost uniform keys with comparative low BER (Bit Error Rate). But Eve may have partial knowledge about the keys by eavesdropping or other ways.

Therefore, in order to gain the absolutely security keys, we must ensure the keys are privacy amplified. Privacy amplification (PA) [4] is a technology, through a public channel, to improve the information confidentiality. Privacy amplification converts the weak secret string W to a uniform key that is fully secret from Eve.

* Corresponding author.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 453–458, 2013.

Privacy amplification technology typically applies a random and public hash function to shorten the weak secret key and reduce the amount of information obtained by Eve as much as possible. By sacrificing partially key information of Alice and Bob，privacy amplification makes the knowledge obtained by Eve been meaningless.

Though the majority researches (such as [5, 6]) about privacy amplification focusing on the theoretical study and proof of security, implementing an efficient privacy amplification scheme in QKD system has been more and more significant.

In this paper, we propose a real-time privacy amplification scheme (RTPA) and implement RTPA in CLIP system [7], which is connected to the quantum communication system. After extensive experiments, the performance and the detail analysis are described in Section III. Experimental results show the efficiency of our proposed RTPA scheme for generating unconditional security keys in quantum cryptography.

2 The Proposed RTPA Scheme

2.1 Privacy Amplification Protocol

After analyzed and researched the classical and quantum privacy amplification theoretical study in [4, 8, 9, 10] and etc., we approach the RTPA protocol (Real-time Privacy Amplification Protocol).

We assume that the key information of Alice and Bob after information reconciliation is W and its length is N, the length of key information used for reconciliation, confirmation and etc. is K, the length of key information may obtained by Eve is T, the security parameter is S, and the final key length is R. We describe the RTPA protocol as follows:

- Alice and Bob select the security parameter S, according to the quantum key state, key length N and other information;
- Alice generates the description information about hash function randomly, the seed string $Seed$ and the shift string $Shift$. $Seed$ and $Shift$ send to Bob through the public channel;
- Alice and Bob construct the hash function f, $f \in F$, $F:\{0,1\}^N \to \{0,1\}^R$, $R = N - T - K - S$;
- Alice and Bob gain the final key y, $y = f(W)$.

In RTPA protocol, hash function f is randomly chosen from class H_3 of universal$_2$. The hash function is described by Toeplitz matrix construction method [11, 12, 13]. After applying the privacy amplification procedure, the final key is unconditionally safe to Eve.

2.2 The RTPA Scheme

The RTPA scheme mainly consists three parts: Hash function construction, data communication and privacy amplification. The architecture of RTPA is shown in Figure 1.

Hash function construction

As shown in Figure 1, the parameter controller carries out the security parameter S and controls the generation of $Shift$ and $Seed$ based on the quantum channel states. Then, the hash function construction module constructs Toeplitz hash function scaling $N \times R$.

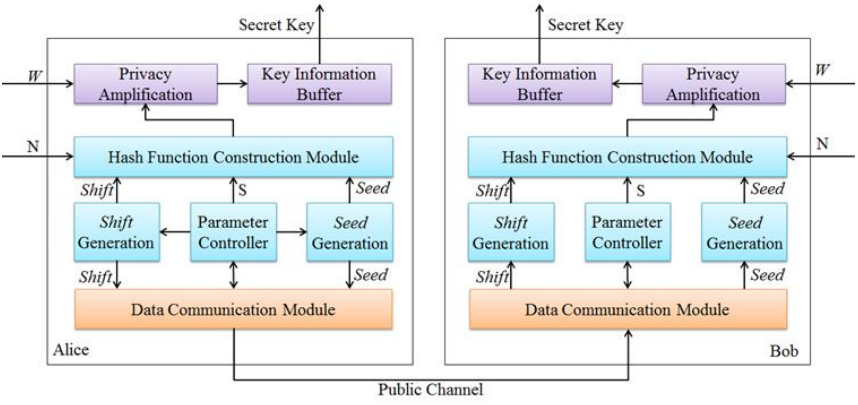


Fig. 1. The Architecture of RTPA

- **Data communication**

  In this part, Alice sends $Shift$ and $Seed$, the description information of hash function , to Bob through a public channel.

- **Privacy amplification**

  Privacy amplification is applied to convert W to an absolutely secret key with length R. These keys used for quantum cryptography are stored in the Key information buffer.

3 Experimental Results and Analysis

The RTPA scheme is implemented in CLIP [7] which is connected to the real quantum communication system. The experimental environment is shown in Figure 2.

We conducted extensive experiments to evaluate the performance of RTPA. We analyzed the privacy amplification overhead, average bit error rate (avBER) of key information.

3.1 Privacy Amplification Overhead

Various hash function constructed for different input key lengths, will lead to different time overhead per privacy amplification process. While the input key length should be long enough in order to gain an absolutely security key, the privacy amplification overhead will be very high. In this experiment, we test the privacy amplification overhead of different hash function scale. The result is shown in Figure 3.

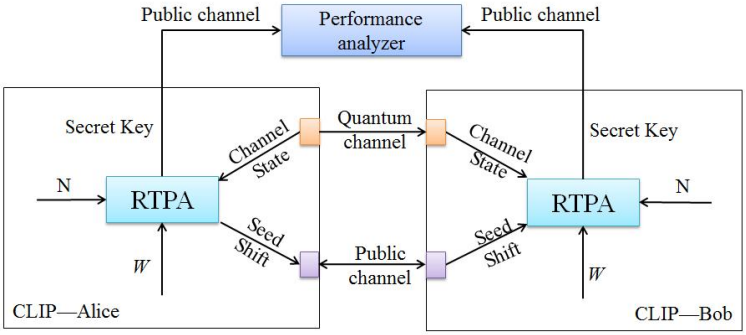


Fig. 2. The experimental environment of RTPA

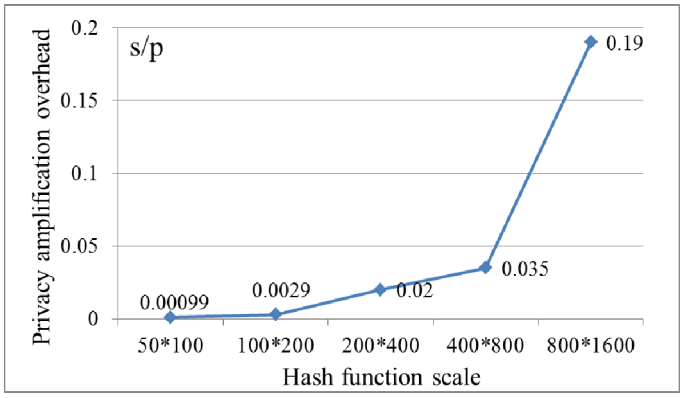


Fig. 3. The privacy amplification overhead of various hash function scales

For example, it will cost 0.19s/p (second per amplification process) when converting a key from 1600 bits to 800bits. Though costing 9.5 times overhead than the scale of 200*400, the security of keys is enhanced by thousands of times. The hash function scale should be balanced between the security demands and the time overhead.

3.2 Average Bit Error Rate

After the procedure of information reconciliation, Alice and Bob have own almost uniform keys with comparative low BER (Bit Error Rate). When applying hash

function to these keys, it may generate quite different strings for Alice and Bob. Therefore, we test the average Bit Error Rate for the final keys with different quantum channel Bit Error Rates.

As it shown in Figure 4, privacy amplification can work effectively when the BER of quantum channel is lower than 0.06. When the BER of quantum channel ranges from 0.06 to 0.10, the information reconciliation procedure still works effectively, the BER after information reconciliation is close to zero, but the BER after privacy amplification is very high. And it doesn't meaningless when the quantum channel BER is higher than 0.10.

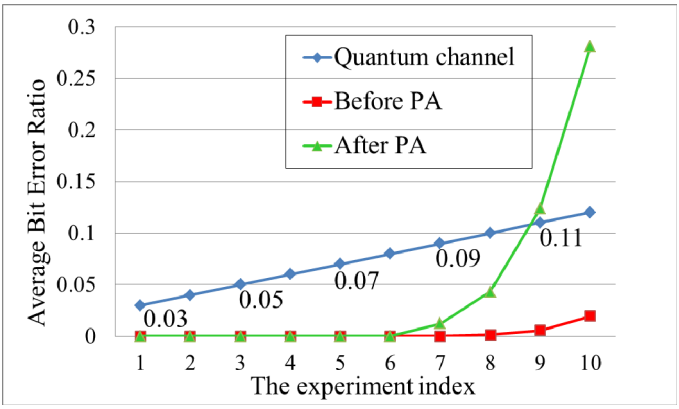


Fig. 4. The Average Bit Error Ratios with different scenes

4 Conclusion

In this paper, we approached the privacy amplification protocol and proposed the RTPA scheme, a real-time quantum privacy amplification procedure in QKD systems. To evaluate the performance of RTPA, we built a prototype QKD system based on CLIP [7]. Experimental results showed the efficiency of our proposed RTPA scheme when the bit error rate of quantum channel is lower than 0.06. The results showed that the performance of RTPA is greatly affected by the quantum channel BER and the information reconciliation. In order to gain an efficient performance, we must enhance the performance of information reconciliation to gain a low BER of key information before privacy amplification.

Acknowledgment. The work described in this paper is partially supported by the grants of the National Basic Research Program of China (973 project) under Grant No.2009CB320503, 2012CB315906;the National High Technology Research and Development Program("863"Program) of China under Grant No. 2011AA01A103, the project of National Science Foundation of China under grant No. 61070199, 61003301, 61103189, 61103194, 61103182, 61202488; the Research Fund for the Doctoral Program of Higher Education of China uner Grant No. 20124307120032,

and supported by Program for Changjiang Scholars and Innovative Research Team in University of the Ministry of Education("Network Technology",NUDT), the Innovative Research Team in University of Hunan Province("Network Technology",NUDT), and the Innovative Research Team of Hunan Provincial natural science Foundation(11JJ7003).

References

1. Bennett, C.H., Brassard, G.: Quantum Cryptography: Public Key Distribution and Coin Tossing. In: Proc. IEEE Int. Conf. Comput. Syst. Signal Process. pp. 175–179 (1984) (QKD)
2. Ekert, A.K.: Quantum cryptography based on Bell theorem. Phys. Rev. Lett. 67, 661–663 (1991) (QKD)
3. Brassard, G., Salvail, L.: Secret Key Reconciliation by Public Discussion. In: Helleseth, T. (ed.) EUROCRYPT 1993. LNCS, vol. 765, pp. 410–423. Springer, Heidelberg (1994)
4. Bennett, C.H., Brassard, G., Crépeau, C., Maurer, U.: Generalized privacy amplification. IEEE Transactions on Information Theory 41(6), 1915–1923 (1995)
5. Watanabe, Y.: Privacy amplification for quantum key distribution. J. Phys. A: Math. Theor. 40, F99–F104 (2007)
6. Renner, R., König, R.: Universally Composable Privacy Amplification Against Quantum Adversaries. In: Kilian, J. (ed.) TCC 2005. LNCS, vol. 3378, pp. 407–425. Springer, Heidelberg (2005)
7. Zou, D., Zhao, B., Wu, C., Liu, B., Yu, W., Ma, X., Zou, H.: CLIP: A Distributed Emulation Platform for Research on Information Reconciliation. In: NBiS 2012, pp. 721–726 (2012)
8. Chandran, N., Kanukurthi, B., Ostrovsky, R., Reyzin, L.: Privacy amplification with asymptotically optimal entropy loss. In: Proceedings of the 42nd Annual ACM Symposium on Theory of Computing, pp. 785–794 (2010)
9. Dodis, Y., Wichs, D.: Non-malleable extractors and symmetric key cryptography from weak secrets. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing, pp. 601–610 (2009)
10. Horváth, T., Kish, L.B., Scheuer, J.: Effective privacy amplification for secure classical communications. EPL 94(2), 28002–28007(6) (2011)
11. Mansour, Y., Nisan, N., Tiwari, P.: The computational complexity of universal hashing. Theoret. Comput. Sci. 107, 121–133 (1993)
12. Krawczyk, H.: LFSR-Based Hashing and Authentication. In: Desmedt, Y.G. (ed.) CRYPTO 1994. LNCS, vol. 839, pp. 129–139. Springer, Heidelberg (1994)
13. Fung, C.-H.F., Ma, X., Chau, H.F.: Practical issues in quantum-key-distribution postprocessing. Phys. Rev. A 81, 012318 (2010)

CSP-Based General Detection Model of Network Covert Storage Channels

Hui Zhu[1,2], Tingting Liu[1], Guanghui Wei[1], Beishui Liu[1], and Hui Li[1]

[1] State Key Laboratory of Integrated Service Networks,
Xidian University, Xi'an, China
xdzhuhui@gmail.com
[2] Network and Data Security Key Laboratory of Sichuan Province

Abstract. A network covert channel is a malicious conversation mechanism, which brings serious security threat to security-sensitive systems and is usually difficult to be detected. Data are hidden in the header fields of protocols in network covert storage channels. In this paper, a general detection model based on formal protocol analysis for identifying possible header fields in network protocols that may be used as covert storage channels is proposed. The protocol is modeled utilizing the Communication Sequential Processes (CSP), in which a modified property of header fields is defined and the header fields are classified into three types in accordance to the extent to which their content can be altered without impairing the communication. At last, verification of the model in Transmission Control Protocol (TCP) shows that the proposed method is effective and feasible.

Keywords: Security modeling, Protocol analysis, Network covert storage channels, Detection, CSP.

1 Introduction and Related Work

A network covert channel is a malicious communication mechanism which can be utilized by attackers to convey information covertly in a manner that violates the system's security policy [1]. The channel is usually difficult to detect and brings serious security threat to security-sensitive systems. Consequently, there is an increasing concern on network covert channels.

There are two types of network covert channels: storage and timing channels.

With the widespread diffusion of networks, many methods have been studied by attackers for constructing network covert channels using a variety of protocols including TCP, IP, HTTP and ICMP [2-4]. For example, Abad proposed [5] an IP checksum covert channels which use the hash collision. Fisk [6] proposed to use the RST flag in TCP and the payload of ICMP protocols to transfer covert message. These covert channel implementations are based on common network or application layer internet protocols. Castiglione et al presented an asynchronous covert channel scheme through using spam e-mails [7]. Moreover, Fiore et al introduced a framework named Selective

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 459–468, 2013.

Redundancy Removal (SRR) for hiding data [8]. It is easy to see that the network covert channels are all based on the various protocols.

More attention has been placed on network covert channels detection. Tumoian et al used the neural network to detect passive covert channels in TCP/IP traffic [9]. Zhai et al [10] proposed a covert channel detection method based on the TCP Markov model for different application scenarios. Gianvecchio et al [11] proposed an entropy-based approach to detect various covert timing channels. The above methods are based either on the anomaly data or on unusual traffic patterns in practical network traffic. It induces that they are hard to find those potential and unknown covert channel vulnerabilities. In this paper, we establish a CSP-based general model to analyze and detect the potential exploits of protocols from the perspective of original design of protocols.

The remainder of the paper is organized as follows. Section 2 introduces the basic concepts of CSP. In Section 3, we give the CSP-based general detection model including the details of establishing and detection steps. In section 4, we test our model in Transmission Control Protocol (TCP). The conclusions are in section 5.

2 CSP (Communicating Sequential Processes)

In CSP [12-15], systems are described in terms of processes which are composed of instantaneous and atomic discrete events. The relations between processes and operations on processes are formalized with operational semantics of the algebra. Many operations on processes and their interrelationships are defined within algebraic semantics of CSP as following:

- $a \to P$ (Prefixing): The process will communicate the event a and then behave as process P.
- $P \square Q$ (Deterministic Choice): This process can behave either as P or Q, but the environment decides on which process to run.
- $a{:}\,\mathrm{A} \to P(a)$ (Prefix Choice): This represents a deterministic choice between the events of the set A which may be finite or infinite. This notation allows representing input and output from channels.
- $c?\,x{:}\,A \to P(x)$: The input can accept any input x of type A along channel c, following which it behaves as $P(x)$.
- $c!\,v \to P$: The output $c!v \to P$ is initially able to perform only the output of v on channel c, then it behaves as P.
- $P \underset{A}{||} Q$ *(Parallel Composition):*Let A be a set of events, then the process behaves as P and Q acting concurrently, with synchronizing on any event in the synchronization set A. Events not in A may be performed by either of the processes independent of the other.
- **Definition 1: trace:** The trace is a model which characterizes a process P by its set traces (P): finite sequences of events which it can perform in a certain period of time.

- **Definition 2: refinement:** A relationship between two CSP processes: trace model refinement. A process P is trace-refined by a process Q, that is $P \sqsubseteq_T Q$, and only if traces (Q) $\subseteq$ traces (P).

3 The CSP-Based General Detection Model

3.1 The General Model Framework

We propose a general detection model focus on the original design details of protocols. The model is used to analyze and detect the covert storage channels vulnerabilities in various layers of protocols. The CSP framework diagram is shown in Figure 1, the model framework includes 5 steps as follows.

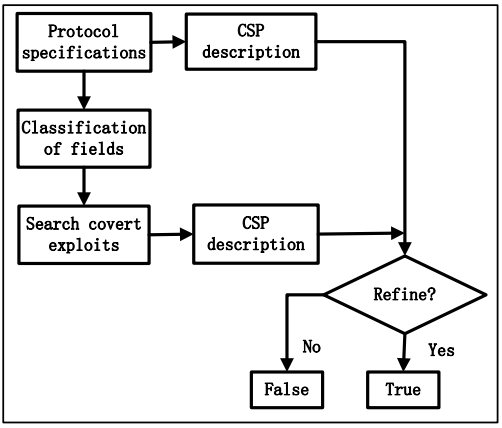


Fig. 1. The CSP-based general model framework

Step1: The original design specifications of protocols and the communication procedure of the protocol interacting entity are analyzed, and then a CSP- based process is established.

Step2: The header fields of protocols are classified into three types—Secure fields, exploited fields-I, exploited fields-II. The details of the classification can be obtained in section 3.2.

Step3: Based on the classification of header fields and the status of the protocol interaction system, a CSP-based search process is established to search for the covert storage channels exploits in the header fields of protocols.

Step4: Based on the hypothesis of the network covert storage channel and vulnerability which have been found in Step3, a CSP-based process of network covert storage channels is established.

Step5: The traces (**Definition1**) of the established processes in Step1 and Step4 are detected at last. It is necessary to find whether the two traces can satisfy the refinement

(**Definition2**) relationship. If the refinement (**Definition2**) relationship can be satisfied by the traces, then there are covert storage channels vulnerabilities in the header fields of protocols, vice verse.

The CSP-based general detection model reduces the existence problem of the covert channels exploits in protocol to the question whether the CSP description of covert storage channels is a refinement of the protocol specifications, which simply the detection of the network covert storage channels.

3.2 Classification of Header Fields

A property named modified property is defined for every header field. The header fields of protocol can be classified into three types according to the modified property as follows:

- **Secure fields:** This type of fields cannot be modified arbitrarily due to their modifications will impair the normal communications. So these fields are secure. For example, the source port field and the destination field of TCP header are secure fields. Once they are modified, the TCP connection cannot be set up.
- **Exploited fields-I:** These fields can be modified arbitrarily with its own merits of making no sense on the normal communication. Such as the reserved fields of TCP headers which are designed for future protocol improvements and the optional header fields of IP.
- **Exploited fields-II:** These fields are needed to guarantee the normal communication under some conditions, so the modifications of them have some restrictions. For example, the TCP urgent point field can be modified to convey messages when the urgent flag of the control bits field is not set.

4 The Verification of Model

4.1 The CSP Description of TCP Connection

In this section, we test and verify the effectiveness and feasibility of the general model in TCP. In order to simplify the analysis and description of TCP, we assume that our TCP interaction system consists of two hosts A and B which are only equipped with an application layer, a TCP entity and two channels between the layers as shown in Figure 2.

The model of TCP describes the connection between a client and a server, and the TCP protocol state machine has six states. We regard host A as a client, and B as a server. Let *Tstate* indicates the set of states. We assume there is a set of different packet types named *PacketTypes*.

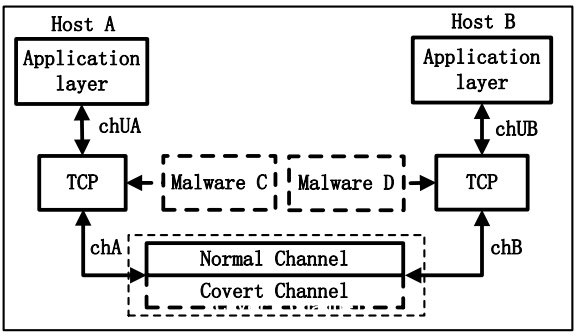


Fig. 2. TCP protocol interaction system

Tstate= {CLOSED, LISTEN, SYN-SENT, SYN-RECEIVED, ESTABLISHED, FIN_WAIT1}
PacketTypes={packet.{syn},packet.{syn_ack},packet.{ack},packet.{rst}, packet.{fin}}

Host A:
TCP_A_SM=
Tstate==CLOSED &
(chUA!!command→
if(command==connect.B)
then (chA!!packet.{syn}→
Tstate:=SYN_SENT))
else（chUA!!Error）→TCP_A_SM
□Tstate==SYN_SENT &
(chA??packet→
If((packet.SYN==1) ∧（packet.ACK==1）
∧（packet. ackNo==seqNo+1））
then (chA!!packet.{ack}→
Tstate:=ESTABLISHED)) → TCP_A_SM
□Tstate==ESTABLISHED &
(chUA??packet→(if(packet.FIN==1)
then (Tstate:=FIN_WAIT1;chA!!packet)
else (chA!!packet)) →TCP_A_SM
|(chA??packet→(if(packet.FIN==1)
then (Tstate:=FIN_WAIT2;chUA!!packet)
else (chUA!!packet) →TCP_A_SM
□Tstate==FIN_WAIT1 &
(chA??packet→(if(packet.FIN==1)
then(Tstate:=FIN_WAIT2)→TCP_A_SM
□Tstate==FIN_WAINT2 &
(wait for close)→TCP_A_SM

Host B:
TCP_B_SM=
Tstate==LISTEN &
(chB??packet→
（if(packet.SYN==1)
then (chB!!packet.{syn_ack}→
Tstate:=SYN_RCVD)
else(discard.packet)) →TCP_B_SM
□Tstate==SYN_RCVD &
(chB?? packet→
if(packet.SYN==1) ∧（packet.ACK==1）
∧（packet.ackNo==seqNo+1）
then(Tstate:=ESTABLISHED)
else (discard.msg)) →TCP_B_SM
□Tstate==ESTABLISHED &
(chuB??packet→(if(packet.FIN==1)
then (Tstate:=FIN_WAIT1;
chB!!packet)
else (chB!!packet)) →TCP_B_SM
□Tstate==FIN_WAIT1 &
(chB??packet→(if(packet.FIN==1)
then(Tstate:=FIN_WAIT2)→TCP_B_SM
□Tstate==FIN_WAIT2 &
(wait for close)
→TCP_B_SM

Fig. 3. The CSP descriptions of TCP in Host A and Host B

Two CSP descriptions for Host A and Host B are shown in Figure 3. The CSP model of the TCP connection is composed of TCP_A_SM and TCP_B_SM.

TCP_CSP=TCP_A_SM ||| TCP_B_SM

4.2 Classification of Header Fields in TCP

According to the rules mentioned in section 3.2, the header fields of TCP are classified into three types as shown in Table 1.

Table 1. The classification of TCP header

type	number	header field		value
Secure fields	1	Source port		0
	2	Destination port		0
	5	TCP header length		0
	7	URG		0
	8	ACK		0
	9	PSH		0
	10	RST		0
	11	SYN		0
	12	FIN		0
	13	Window size		0
exploited fields-I	6	Reserve field(6 bit)		1
exploited fields-II	3	SeqNo	SYN==1,ACK==0	V=1
	4	AckNo	SYN==1,ACK==0	V=1
			ACK==1	V=0
	14	checksum		V=1
	15	Urg_p	URG==0	V=1
			URG==1	V=0

The process C_exploit(X) search for the covert storage channels exploits in TCP according to the rules based on the modified property and the specifications of TCP connection. Figure 4 shows the CSP process C_exploit(X). After searching for the network covert storage channel exploits through the process C_exlpoit(X), we have obtained three results based on the different states of TCP connection.

- When the TCP connection is beginning to establish (Tstate=CLOSED): C_exploit(X)={seqNo, ackNo, Reserve_field, checksum, Urg_p}
- When the TCP entities are not in closed state and the urgent flag of the control bits is not used:
  C_exploit(X)={Reserve_field, checksum, Urg_p}
- When the TCP entities are not in closed state and the urgent flag of the control bits is set:
  C_ exploit(X)={Reserve_field,checksum}

```
C_exploit (X) = learn? Packet:Packet.i→C_ exploit (Judge(X∪{i}))
        □ (if X=∅ then overt→C_ exploit (X)
                                  else  covert→C_ exploit (X))
Subprocess Judge (X∪{i} is as follow:
Judge (X∪{i})=(if(Tstate==CLOSED)∧(cmd.connect.B)→add.seqNo);
                (if(Tstate==SYN-SENT)∧(SYN==1) ∧(ACK==0) →add.ackNo);
                (if(Tstate≠CLOSED) →add.Reserve_field);
                (if(packet. Window_size≠null→add.checksum);
                (if(packet.URG==0)→add.Urg_p)
```

Fig. 4. The CSP process C_exploit(X)

4.3 The CSP Description of Covert Storage Channels

As have been shown in Figure 2, there are two malwares C and D who hide inside the TCP interaction system.

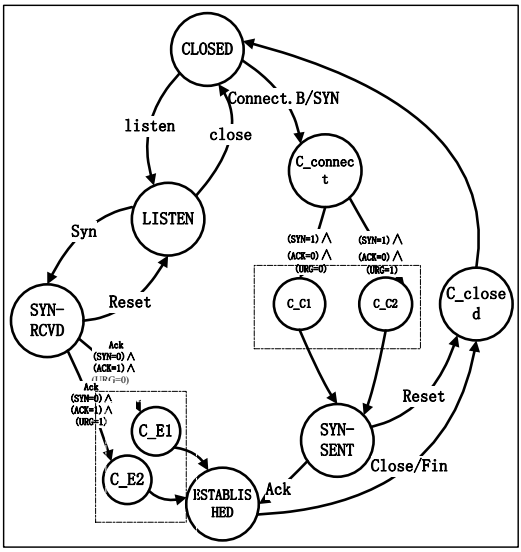


Fig. 5. The finite state diagram of TCP interaction system

Assume that the malwares C and D set up a covert storage channel, of which C is the sender of the covert channel, D is the receiver. Let the set ***Cstate*** indicate the states of covert storage channels. The set *IS_state* indicates the global states of TCP interaction system which include the states of covert channel. Then, *IS_state=Tstate* ∪ *Cstate*. Figure 5 depicts the state transition diagram of the whole TCP interaction system.

Cstate= {C_closed, C_listen, C_connect, C_c1, C_c2, C_e1, C_e2}

The malwares monitor the status of TCP connection and utilizes the exploited fields of TCP to transmit the covert message. The CSP process of malware C is shown in Figure 6. The CSP description of malware D is similar to the malware C, so we omit its description here. The process IS_CSP indicates the interaction system of TCP protocol entities and two malwares.

```
C_Channel=(listen??chUA)→
            (if(command==connect.B)∧(Tstate==CLOSED)
              then(Cstate:=C_connect))→C_Channel_modu(Cstate)
              else if(Tstate==SYN_SENT)then(Cstate:=C_established)
              else if(Tstate==ESTABLISHED)then(Cstate:=C_established)
              else if(Tstate==FIN_WAIT1)then(Cstate:=C_closed)
Subprocess C_Channel_modu(Cstate) is as follow:
C_Channel_modu(Cstate)=
        Cstate=C_connect  &
        (if(SYN==1)∧(ACK==0)∧(URG==0)
        then(Cstate:=C_c1;modu(packet.{syn})→chA!!packet. {syn} .(covert_msg))
        else if((SYN==1)  ∧(ACK==0)∧(URG==1)
        then(Cstate:=C_c2;modu(packet.{syn})→chA!!packet. {syn} .(covert_msg)) )
□Cstate=C_ established  &
  (if(URG==0)
  then (Cstate:=C_e1;modu(packet.{syn_ack})→chA!!packet.{syn_ack}.(covert_msg))
  else (Cstate:=C_e2;modu(packet.{syn_ack})→chA!!packet.{syn_ack}.(covert_msg))
```

Fig. 6. The CSP process of malware C

The description of network covert storage channels is
Covert_channel= C_Channel ||| D_Channel
Then, the following IS_CSP is obtained.
IS_CSP=TCP_CSP ||| Covert_channel
=TCP_A_SM ||| TCP_B_SM ||| C_Channel ||| D_Channel

4.4 Analysis of the Verification

In this section, we analyze and verify the existence of network covert storage channels under the normal communication of TCP using trace model refinement (**Definition 2**). It means that whether the traces of IS_CSP are subsets of traces of TCP_CSP which allow precisely the valid traces.

The CSP model can detect whether the network covert storage channels are possible under the normal communication of TCP. Figure 7 shows an example of traces of TCP_CSP and IS_CSP. As we can see that tr(IS_CSP) is a subset of tr(TCP_CSP):

```
tr(TCP_CSP)=
<cmd.connect.B,send.A.B.packet.{syn},receive.B.A.packet.{syn_ack},send.A.B.packet.{ack},receiv
e.B.A.packet.....>
tr(IS_CSP)=
<cmd.connect.B,send.A.B.packet.{syn}.(seqNo=covert_msg),receive.B.A.packet.{syn_ack}.(urgent
_pointer=covert_msg),...>
```

Fig. 7. An example of traces of TCP_CSP and IS_CSP

$$tr(IS\_CSP) \subseteq tr(TCP\_CSP)$$

From the trace of IS_CSP, we can see that the malware C modulates the initial sequence number to convey the covert message which has been utilized and found by the previous researchers. For example, Rutkowska [16] implemented a network storage channel utilizing the initial sequence number named NUSHU. On basis of the definition 2, we come to the conclusion that the CSP model IS_CSP refines TCP_CSP.

$$TCP\_CSP \sqsubseteq_T IS\_CSP$$

```
tr(TCP_CSP)=
<cmd.connect.B,send.A.B.packet.{syn},receive.B.A.packet.{syn_ack},send.A.B.packet.{ack},.....>
tr(IS_CSP)=
<cmd.connect.B,send.A.B.packet.{syn}.(reserved_fields=covert_msg),receive.B.A.packet.{syn_ack}.(reser-
ved_fields=covert_msg),....>
```

Fig. 8. Another example of traces of TCP_CSP and IS_CSP

Figure 8 depicts other similar traces as well. The traces show us that the malwares C and D utilize the reversed field to convey covert message. This covert storage channel have been found and studied by Handel [17].

The verification of TCP proves that the CSP-based general model might yield several very similar covert storage channels in other network protocols in terms of the modified property of header fields and the specifications of network protocols.

5 Conclusion

In this paper, we propose a CSP-based general detection model for analyzing and detecting the network covert storage channels in network protocols. In our model, we describe the protocol interaction system based on the original design specifications of protocols. Besides, we define a modified property for every header field, and classify the header fields into three types based on this property. We establish a search process for searching the potential covert exploits in the header fields of protocols. Then we establish a network covert storage channel based on the hypothesis of network covert storage channels, and verify the covert channels based on trace refinement. Finally, the model of this CSP formal method is illustrated and verified in

TCP. The result of the verification shows that the general model is effective and feasible in finding the covert storage channels.

The CSP-based general detection model is modular, so it can be easily extended to describe the other network protocols and detect the covert channels hidden in them. In the future, we will try to establish a formalized method for detecting and analyzing the covert timing channels.

References

1. Snoeren, A., Partridge, C., Sanchez, L.: Single Packet IP Trace back. ACM/IEEE Transaction on networking 10(6), 721–734 (2002)
2. Zander, S., Armitage, G., Branch, P.: A Survey of Covert Channels and Countermeasures in Computer Network Protocols. IEEE Communications Surveys and Tutorials 9(3), 44–57 (2007)
3. Ahsan, K., Kundur, D.: Practical Data Hiding in TCP/IP. In: ACM WKSP Multimedia, 7–14 (2002)
4. Cauich, E., Gardenas, R.G., Watanabe, R.: Data Hiding in Identification and Offset IP Fields. In: 5th International Symposium (2005)
5. Abad, C.: IP checksum covert channels and selected hash collision. Technical report (2001)
6. Fisk, G., Fisk, M., Papadopoulos, C., Neil, J.: Eliminating Steganography in Internet Traffic with Active Wardens. In: Petitcolas, F.A.P. (ed.) IH 2002. LNCS, vol. 2578, pp. 18–35. Springer, Heidelberg (2003)
7. Castiglione, A., Santis, A.D., Fiore, U., Palmieri, F.: An asynchronous covert channel using spam. Computers and Mathematics with Applications 63(2), 437–447 (2012)
8. Fiore, U.: Selective Redundancy Removal: A Framework for Data Hiding. Future Internet 2(1), 30–40 (2010)
9. Tumoian, E., Anikeev, M.: Network based detection of passive covert channels in TCP/IP. In: Proceedings of the IEEE Conference on Local Computer Networks 30th Anniversary, pp. 802–809 (2005)
10. Zhai, J., Liu, G., Dai, Y.: A covert channel detection algorithm based on TCP Markov model. In: Proceedings of Second International Conference on Multimedia Information Networking and Security, pp. 893–897 (2010)
11. Gianvecchio, S., Wang, H.: An Entropy-Based Approach to Detecting Covert Timing Channels. IEEE Transactions on dependentable and secure computing 8(6), 785–797 (2011)
12. Hoare, C.A.R.: Communicating Sequential Processes. In: Communications of the ACM, pp. 666–677 (1978)
13. Brookes, S.D., Hoare, C.A.R., Roscoe, A.W.: A theory of Communicating Sequential Processes. Journal of the ACM 31, 560–599 (1984)
14. Roscoe, A.W.: The theory and practice of concurrency, s. l. Prentice Hall (1998)
15. Schneider, S.A.: Concurrent and real-time systems: the CSP approach, s. l. John Wiley (1999)
16. Rutkowska, J.: The implementation of passive covert channels in the Linux kernel, `http://invisiblethings.org/papers.html`
17. Handel, T.G., Sandford, M.T.: Hiding Data in the OSI Network Model. In: Anderson, R. (ed.) IH 1996. LNCS, vol. 1174, pp. 23–38. Springer, Heidelberg (1996)

Trustworthy Opportunistic Access to the Internet of Services

Alessandro Armando[1], Aniello Castiglione[2], Gabriele Costa[1], Ugo Fiore[3], Alessio Merlo[1,4], Luca Verderame[1], and Ilsun You[5]

[1] Università degli Studi di Genova, Genova, Italy
name.surname@unige.it
[2] Università degli Studi di Salerno, Fisciano, Italy
castiglione@ieee.org
[3] Università di Napoli Federico II, Napoli, Italy
ufiore@unina.it
[4] Università E-Campus, Novedrate, Italy
alessio.merlo@uniecampus.it
[5] Korean Bible University, Seoul, Republic of Korea
isyou@bible.ac.kr

Abstract. Nowadays web services pervade the network experience of the users. Indeed, most of our activities over the internet consist in accessing remote services and interact with them. Clearly, this can happen only when two elements are available: (*i*) a compatible device and (*ii*) a suitable network connection. The recent improvement of the computational capabilities of mobile devices, e.g., tablets and smartphones, seriously mitigated the first aspect. Instead, the inappropriateness, or even the absence, of connectivity is still a major issue.

Although mobile, third generation (3G) networks can provide basic connectivity, complex interactions with web services often require different levels of Quality of Service (QoS). Also, 3G connectivity is only available in certain areas, e.g., user's country, and purchasing temporary connection abroad can be very costly. These costs weigh down on the original service price, seriously impacting the web service business model.

In this paper we describe the problems arising when considering the orchestration of service-oriented opportunistic networks and we present the assumptions that we want to consider in our context. We claim that our model is realistic mainly for two reasons: (*i*) we consider state-of-the-art technology and technical trends and (*ii*) we refer to a concrete problem for service providers.

1 Introduction

The evolution of Web 2.0 as well as the spread of cloud computing platforms have pushed customers to use always more remote services (hosted in a cloud or a server farm) rather than local ones (installed on personal devices). Such paradigm shift has basically improved the role of the network connectivity. Indeed, the access to remote services as well as the user experience, strongly depends on the network availability and the related performances (i.e., QoS).

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 469–478, 2013.

To get evidence of this, let us consider a set of cloud users travelling through an airport and needing to access remote services from their device (e.g. smartphone, tablets, laptop) to complete their job. Presently, telecommunications companies sell internet connection for fixed time slots inside the airport, by means of 3G or wireless connections. Thus, each of these cloud users is compelled to subscribe, individually, to such internet connections, thereby getting extra charge to access the remote service. Moreover, users often do not get through the purchased connection, using less bandwidth or disconnecting before the end of the time slot. Then, such scenario leads to a non-negligible waste of purchased resources and money that may be reduced whether proper architectural or software solutions would allow, for instance, cooperation and resource sharing among the cloud users.

In this paper, we cope with such problem by investigating the adoption of the Software Defined Networking (SDN) paradigm as potential solution to build and manage QoS-constrained and on demand connection among mobile devices. In particular, we describe the main issues arising when trying to orchestrate a group of mobile devices that participate in an opportunistic network. Besides the difficulty of finding valid orchestration, e.g., in terms of QoS, we also present the security concerns at both network and device level. Finally we introduce a case study illustrating how our assumptions apply to a real life web service.

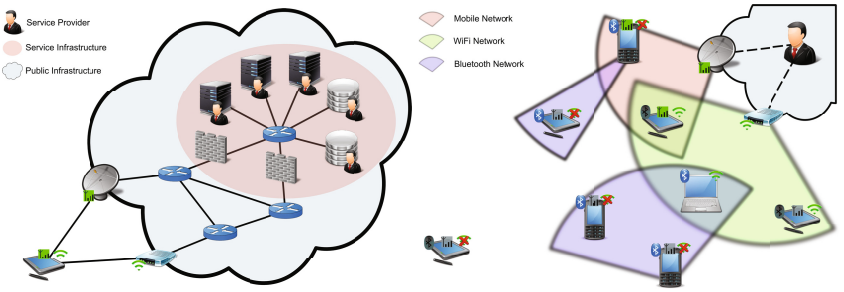
This Paper Is Structured as Follows. In Section 2 we state the problem of orchestrating opportunistic, service-oriented networks. Then, Section 3 describes the main security issues arising in this context and how to deal with them. In Section 4 we present our case study and its features. Finally, in Section 5 we survey on some related works and Section 6 concludes the paper.

2 Problem Statement

A provider P of a service S relies on a network infrastructure implementing S. The implementation of S is designed to meet both functional, e.g., accessibility, QoS and responsiveness, and non functional, e.g., security and fault tolerance, requirements. Moreover, through proper testing procedures, evidences that the implementation of S complies with these requirements have been produced and collected by P. In order to access S, customers need a network enabled device, e.g., laptops, tablets and smartphones, that can connect to and interact with S (typically by means of a client application). This scenario is depicted in Figure 1a.

Clearly, when a suitable connection is not available, the customer has no access to S. In order to access S, customers might enable a new connection, e.g., by buying a (costly and slow) 3G or a (local) wifi connection from a connectivity provider. This approach requires an existing infrastructure to be present and, definitely, charges extra costs on customers that, possibly, already pay for S.

Recent technological trends have highlighted that mobile devices can share their connectivity by playing the role of an access point. This technology, known as *tethering*, exploits multiple connection paradigms, e.g., wifi, bluetooth and



(a) Standard access configuration. (b) Service unreachable to customer.

Fig. 1. Mobile access to a web service

IR, to create local networks. For instance, a mobile device can use wifi tethering to share its 3G connection with a group of neighbors. Although a single device has serious limitations, e.g., battery consumption, computational overhead and bandwidth exhaustion, populated areas are typically characterized by the presence of many devices. Thus, a proper orchestration of (a group of) these devices can lead to more powerful and reliable networks.

2.1 Local Area Configuration

A customer C having no network connectivity is logically isolated from service S. However, the customer's device is physically surrounded by networked devices. These devices connect to one or more networks by means of different channels. Schematically, an instance of such a configuration is reported in Figure 1b.

Mobile Agents. Mobile devices have heterogeneous hardware profiles, e.g., memory, computational power, presence/absence of bluetooth, etc. Also, their configuration can change over time, e.g., on device battery charge and position.

In general, we can consider each device to be a computational platform that can install and run (at least small) applications. Moreover, we assume that all the (enabled) devices run software supporting basic orchestration steps.

Communication Protocols. Connected devices use different channels, e.g., bluetooth and wifi, to establish connections. These channels have different features and, in general, have been designed for different purposes. We call a *pit* a device having direct access to the internet. Hence, the devices must create a network where one or more pits are present. Other resources can be present in the network, e.g., computation and memory nodes, and they can be exploited for the service delivery.

Device Contracts. Each device holds a precise description of its features and requirements, namely a *contract*. Contracts describe which kind of activities can be carried out by the device. Examples of entries of a contract are:

- Available disk space, i.e., the amount of memory that the device can offer.
- Available connections, i.e., channel types, bandwidth, etc.
- Computational capacity, i.e., whether the device can carry out computation.

Each feature can be associated to a precise cost that must be paid to access/use it. Informally, we can see a contract as a list of predicates like:

NET. Internet: 3G (Bandwidth: 3.2 MB/sec; Cost: 0.05 €/MB) + WiFi (Bandwidth: 14.4 MB/sec; Cost: 0.01 €/MB);
LINK. Bluetooth: (Bandwidth: 720 Kb/sec; Cost: 0 €/sec);
DISK. Space: 2 GB; Cost: 0.01 €/MB; Expiration time: 60';
CPU. Speed: 800 MHz; Cost: 0.02 €/sec;

For instance, the first rule says that the device can connect to the internet in two different ways (i.e., 3G and WiFi) and describes the differences in costs and bandwidth. Instead, the meaning of the third clause is that the device can offer up to 2 GB of memory space at the given cost per MB. Also, the contract says that after 60 minutes stored data will be deleted.

Other devices can retrieve a contract and compare it against their requirements. Moreover, when a contract does not satisfy certain requirements, a *negotiation* process is started. Negotiation consists in proposing modifications to the original contract clauses. If the contract owner accepts the proposed corrections, a new, suitable contract is used in place of the previous, unfitting ones.

2.2 Network Orchestration

Network orchestration plays a central role. Indeed devices must cooperate in a proper way in order to achieve the network goals. Among the recent proposals for the organization of networks, *Software Defined Networking* (SDN) is receiving major attention.

Software Defined Networking. The main feature of SDN is a clear distinction between control plane and data plane in network choreographies. Mainly this approach allows for exploiting centralised service logic for the network orchestration. Typically, network nodes take responsibility for both data transfer and network organization activities promiscuously. This behavior is acceptable when networks are composed by dedicated nodes, i.e., platforms (hardware and software) dedicated to the network management. However, under our settings we cannot expect to have homogeneity in nodes configurations. Indeed, nodes configurations may differ for many reasons, e.g., hardware, battery state, user's activities and security policies. Hence, we must expect that the network management is carried out by some dedicated devices in a partially distributed way.

Nodes offering advanced computational capabilities can take responsibility for the control activities. These activities include node orchestration, network monitoring and reaction to changes. Since nodes do not have any pre-installed orchestration software, mobile code must be generated and deployed dynamically. Figure 2 represents this process.

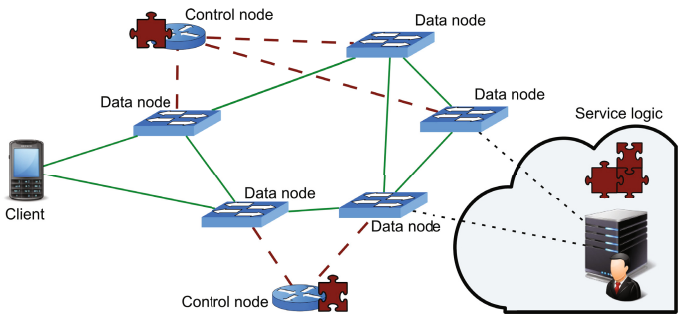


Fig. 2. Control and data nodes participating in an orchestration

Solid lines represent data links, i.e., channels used for service-dependent traffic. Instead, dashed lines denotes control channels, i.e., used for control activities. Control instructions are generated by the service provider being the entity holding the service logic. A control node receives a piece of software (jigsaw piece) that is responsible for managing the behavior of (part of) the network.

3 Security Issues

Many security threats can arise during the recruiting, negotiation and execution phases. All the security aspects can appear at either (*i*) network level or (*ii*) service level. Below, we list and describe the main security issues showing whether they affect the first or second of these layers.

3.1 Network Security

Devices in opportunistic networks build transient and goal-driven networks, thus behaving as peers. Hence, most of the security issues at network level resemble those of P2P networks. By joining an opportunistic network, each device gets potentially unknown neighbours and exchange data with them. In this context, confidentiality and integrity are major guarantees to provide to the final user, since information sent to the service S may be corrupted or intercepted by malicious devices. Device authentication is also required in order to recognize and isolate malicious devices.

Authenticity. Usually devices are uniquely characterized, e.g., by the MAC address or IMEI code. However, a strong authentication relating a device with a physical user is hard to achieve at this level. Also a global authentication in a network is hardly achievable, due to the lack of a central authority and the heterogeneity of device platforms. However, mutual and pairwise authentication between devices may be easily carried out. In this context, from the single-device point of view the authentication is aimed at 1) allowing honest devices to

recognize and isolate malicious ones, and 2) building temporary trust relationship between trusted and authenticated devices in order to share bandwidth, memory and disk resources. To meet these targets, the adoption of gossiping algorithms [6], combined with cooperative access control mechanisms [12] can be adopted.

Confidentiality. Message confidentiality is a main concern since a device reaches the service by sending data through unknown and untrusted devices, without the possibility to trace its own traffic. However, confidentiality at this layer can be granted by the use of secure channels built at higher layer. For instance, HTTPS channels established between the source device and the service are suffice to provide the required confidentiality through traffic encryption. Secrecy provided by HTTPS channels is not easily breakable, in particular for single devices in the networks.

Integrity. Ciphering data grants secrecy but does not prevent devices from tampering the traffic they receive. Thus, the use of integrity scheme can be envisaged in opportunistic networks. There exist integrity schemes based on shared keys and private/public key. The choice between shared key (e.g., MAC schemes [14]) and public/private key schemes (e.g., DS schemes [4] and Batch verification schemes [10]) depends on the contingency of the opportunistic network. Besides, the use of such schemes requires, at most, the installation of simple libraries or programs on the device.

3.2 Service Level Security

Here we can identify two groups of entities aiming at different security goals: (*i*) the service-customer coalitions and (*ii*) the control-data nodes.

Service-Customer Security. The service provider and its customers share a common goal, i.e., enabling the customer to access the service according to a given SLA. Among the clauses of the agreement, security policies prescribes how the service handles the customer's data and resources. In general, the provider can rely on a trusted service infrastructure. However, in our scenario the service is delivered by a group of, potentially untrusted, devices which extend the trusted infrastructure. Intruders could join the network and perform some disrupting attack, e.g., denial of service, also depending on the service typology.

On the other hand, the service can include security instructions in the code deployed on control nodes. In this way, control nodes can monitor the behavior of (a portion of) the network. Monitoring allows control nodes to detect intruders and, possibly, cut them off. Even more dangerously, the intruder could be a control node. In this case, the service can detect the misbehaving control node by directly monitoring its behavior. Control node monitoring can rely on other control nodes, including the (trusted) customer. Hence, a group of control nodes can isolate a malicious peer when detected. Still, control nodes collusion represent a major threat and mitigation techniques could be in order.

Nodes Security. Data and control nodes have a different security goal. Since they join an orchestration upon contract acceptance, their main concern is about avoiding contract violations. Being only responsible for packets transmission, data nodes can directly enforce their contract via resources usage constraints.

Control nodes require more attention. As a matter of fact, they receive orchestration code from the server and execute it. The execution of malicious software can lead to disruptive access and usage of the resources of the device. Thus, a control node must have strong, possibly formal, guarantees that the mobile code is harmless, i.e., it respects the given contract.

A possible solution consists in running the received code together with a security monitor. Security monitors follow the execution of a program and, when they observe an illegal behavior, run a reaction procedure, e.g., they can throw a security exception. A security monitor comparing the mobile code execution against a given contract is an effective way to ensure that no violations take place. Although monitoring requires extra computational effort, lightweight implementations causing small overheads have been proposed, e.g., see [8].

Another approach exploits static verification on to avoid run-time checks. Briefly, the service provider can run formal verification tools, e.g., a model checker, before delivering the mobile code. The model checker verifies whether the code satisfies a given specification, i.e., the contract, and, if it is not the case, returns an example of a contract violation. Instead, if the process succeeds, a proof of compliance is generated. Using *proof-carrying code* [13] the proof is then attached to the code and, then, efficiently verified by the control node. A valid proof ensure that the code execution cannot violate the contract of the node.

4 Meeting at the Airport: A Case Study

We consider the following scenario. A e-meeting service offers to its customers the possibility to organise and attend to virtual meetings. A meeting consists of a group of customers that use (*i*) a VoIP system for many-to-many conversations and (*ii*) file sharing for concurrently reading and writing documents.

Private companies buy annual licenses. Then, employees install a free client application on their devices and access the service using proper, company-provided credentials. Nevertheless, company employees use to travel frequently and, often, need to buy wireless access in airports and train stations. This practice causes extra, variable charges on the service usage.

Service Requirements. The two service components, i.e., VoIP and file sharing, have different features. Mainly, the VoIP service has precise constraints on transmission times in order to make the communication effective. In order to respect this constraint, the service can reduce the voice encoding (up to a minimal threshold) quality whenever slow connections risk to cause delays.

Instead, the file sharing system must guarantee that documents are managed properly. Roughly, users can acquire the ownership of a document and modify it until they decide to release the control. Each time a document is retrieved

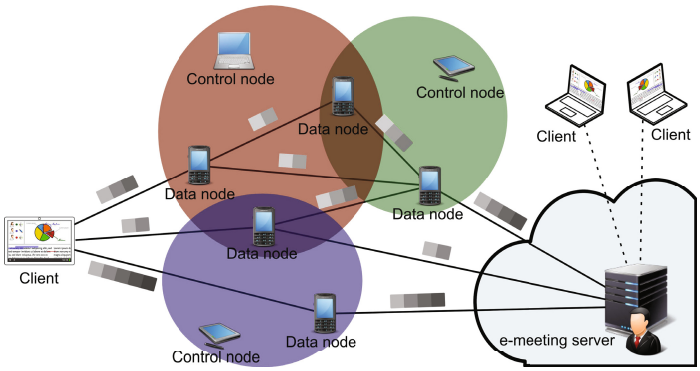


Fig. 3. Orchestration providing opportunistic access to e-meeting

(submitted) it is downloaded from (uploaded to) a network repository database. Document loading and saving are not time critical operations, i.e., they can be delayed, but data consistency must be guaranteed.

Network Structure. In order to set up a suitable network, the client starts a *recruiting* procedure. Briefly, it floods with a request message its neighbours (up to a fixed hops number) and collects their contracts. If the set of received contracts satisfies preliminary conditions, e.g., sufficient nodes density ad existence of internet-enabled nodes, the negotiation process starts.

Negotiation requires interaction with the web service. To do this, at least one of the nodes having internet access must take responsibility for sending the negotiation information to the orchestration service. This information includes nodes contracts and topology description, i.e., nodes neighbours tables. The orchestrator check whether the network configuration satisfies minimal requirements and returns contract proposals for the control nodes. The nodes receive the negotiated contracts and decide whether to accept or reject it. If a contract is rejected, the process can be repeated[1]. When all the control nodes accept the proposed contracts the service send them a piece of software implementing part of the distributed orchestration algorithm. Each node verifies the validity of the received code and starts the orchestration procedure. The resulting network organization is depicted in the figure below.

Intuitively, each control node is responsible for coordinating the activities of a group of data nodes (rounded areas). Data nodes are responsible for transmitting network traffic and they are recruited and managed by control nodes. Also, control nodes must react to a plethora of possible events, e.g., topology changes, data and control node fall or performances decay.

[1] We assume that nodes cannot reject a contract respecting its own original clauses. For instance a node offering 2 GB of disk space can reject a request for 3 GB but not one for 1 GB.

5 Related Work

Many technologies are related to our model. Here we briefly describe those that, in our view, better apply to this context.

Just recently, software defined networking received major attention. Among the others, OpenFlow [11,9,15] is the leading proposal of an implementation of SDN. Basically, OpenFlow allows network managers to programmatically modify the behavior of networks. Although it is currently applied to standard network infrastructure, i.e., standard routers and switches, this technology seems to be also suitable for mobile devices. Hence, we consider it to be a promising candidate for the implementation of orchestration tools.

Formal verification plays a central role under our assumptions and it appears at several stages. Mainly, contract-based agreements require formal analysis techniques for granting that implementations satisfy a contract. A standard method for this is model checking [7,2]. However, also proof verification is crucial for allowing network nodes to check the proof validity when the source in an untrusted service. This step can be implemented by using proof-carrying code [13].

Being a main concern, code mobility and composition environment must include proper security support. In particular, policies composition techniques must be included in any proposed framework. Local policies [3] represent a viable direction for allowing several actors to define their own security policies, apply them to a local scope and compose global security rules efficiently. Also, since our proposal is based on mobile devices technology, specific security solutions for mobile OSes must be considered. In this sense, in [1] the problem of securing the Android platform against malicious applications has been studied.

Finally, also dynamic monitoring appear to be necessary for managing and re-organizing the network in case of necessity, e.g., upon failure discovery. A possible solution consists in using the approach presented by Bertolino et al. [5] for retrieving and collecting information about nodes behavior. Instead, for what concerns security monitoring, a possible approach is presented in [8]. Since this proposal has been tested on resource limited devices, it seems a good candidate for avoiding computational loads on network nodes.

6 Conclusion

In this paper, we described the possibility of applying Software Defined Networking (SDN) paradigm as potential solution to build and manage opportunistic connection among mobile devices and web services. In particular, we described the main issues arising when trying to orchestrate devices that share the goal of implementing a QoS compliant network. Also, we considered the security issues deriving from such a model and possible approaches and countermeasures. Finally, we presented a case study that highlights the main aspects that must be considered under our assumptions.

References

1. Armando, A., Costa, G., Merlo, A.: Formal modeling and reasoning about the Android security framework. In: Proc. of 7th International Symposium on Trustworthy Global Computing (2012) (to appear)
2. Baier, C., Katoen, J.-P.: Principles of Model Checking (Representation and Mind Series). The MIT Press (2008)
3. Bartoletti, M., Degano, P., Ferrari, G., Zunino, R.: Local policies for resource usage analysis. ACM Transactions on Programming Languages and Systems, TOPLAS 31(6), 23 (2009)
4. Bellare, M., Rogaway, P.: The Exact Security of Digital Signatures - How to Sign with RSA and Rabin. In: Maurer, U.M. (ed.) EUROCRYPT 1996. LNCS, vol. 1070, pp. 399–416. Springer, Heidelberg (1996)
5. Bertolino, A., Calabrò, A., Lonetti, F., Di Marco, A., Sabetta, A.: Towards a Model-Driven Infrastructure for Runtime Monitoring. In: Troubitsyna, E.A. (ed.) SERENE 2011. LNCS, vol. 6968, pp. 130–144. Springer, Heidelberg (2011)
6. Boyd, S., Ghosh, A., Prabhakar, B., Shah, D.: Gossip algorithms: Design, analysis and applications. In: Proceedings of IEEE INFOCOM, pp. 1653–1664 (2005)
7. Clarke Jr., E.M., Grumberg, O., Peled, D.A.: Model checking. MIT Press, Cambridge (1999)
8. Costa, G., Martinelli, F., Mori, P., Schaefer, C., Walter, T.: Runtime monitoring for next generation java me platform. Computers & Security 29(1), 74–87 (2010)
9. Pfaff, B., et al.: OpenFlow Switch Specification. OpenFlow (February 2011), `http://www.openflow.org/documents/openflow-spec-v1.1.0.pdf`
10. Gasti, P., Merlo, A., Ciaccio, G., Chiola, G.: On the integrity of network coding-based anonymous p2p file sharing networks. In: Proceedings of the 2010 Ninth IEEE International Symposium on Network Computing and Applications, NCA 2010, pp. 192–197. IEEE Computer Society, Washington, DC (2010)
11. McKeown, N., Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, L., Rexford, J., Shenker, S., Turner, J.: OpenFlow: enabling innovation in campus networks. SIGCOMM Comput. Commun. Rev. 38(2), 69–74 (2008)
12. Merlo, A.: Secure cooperative access control on grid. Future Generation Computer Systems 29(2), 497–508 (2013)
13. Necula, G.C.: Proof-carrying code. In: Proceedings of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL 1997, pp. 106–119. ACM, New York (1997)
14. Preneel, B., Van Oorschot, P.C.: On the Security of Two MAC Algorithms. In: Maurer, U.M. (ed.) EUROCRYPT 1996. LNCS, vol. 1070, pp. 19–32. Springer, Heidelberg (1996)
15. Tootoonchian, A., Ganjali, Y.: Hyperflow: a distributed control plane for openflow. In: Proceedings of the 2010 Internet Network Management Conference on Research on Enterprise Networking, pp. 3–9. USENIX Association, Berkeley (2010)

Architecture of Network Environment for High-Risk Security Experimentation

Xiaohui Kuang[1,2], Xiang Li[1,2], and Jinjing Zhao[1,2]

[1] National Key Laboratory of Science and Technology on Information System Security, Beijing
[2] Beijing Institute of System Engineering, China
Xiaohui_kuang@163.com,
{lixiang8358,misszhaojinjing}@hotmail.com

Abstract. Adequate Environment for conducting security experiments and test under controlled, safe, repeatable and as-realistic-as-possible conditions, are a key element for the research and development of adequate security solutions and the training of security personnel and researchers. In this paper, a new large-scale network experimental environment for high-risk security research was put forward. The main idea was using isolated computing clusters to obtain high levels of scale, manageability and safety by heavily leveraging virtualization technology, separating experiment and control network and multilayer sanitization.

Keywords: security experimental environment, architecture, virtualization, characteristic analysis.

1 Introduction

Experimentation is a keystone of scientific method and of technology innovation and development. The success of Internet can be attributed to many factors. However, experimentation environment played an important role. Today more and more researchers involved in designing network services, protocols and security mechanism rely on results from emulation environments to evaluate correctness, performance and scalability [1][2][3][4][5][6]][7].

To better understand the behavior of these applications and to predict their performance when deployed across the Internet, the emulation environments must closely match real network characteristics. But, these emulation environments pay attention to scalability, fidelity and flexibility, don't take into account about risk and sanitization. The high-risk experiments such as network worms, botnets, and viruses and so on, are difficult to conduct in these network emulation environments.

In this paper we describe a new architecture of network environment for conducting high-risk security experimentation. The key idea behind our environment is the use of isolated, special purpose computing clusters, which heavily use virtualization technology, physical and logical separation, and cluster management tools in order to attain a high level of scalability, flexibility, isolation, and sanitization. The main

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 479–484, 2013.

contribution of this paper is 1) to analyze the requirement of high-risk security experiment, 2) to describe the architecture of a large-scale network environment that match the requirement, 3) to analysis the characteristic of the environment, 4) to describe the test on the environment.

The rest of the paper is organized as follows: Section 2 presents the requirement of security research facilities should match. Section 3 describes the architecture of our experimental environment named LNESET (Large-scale Network Emulation environment for high-risk Security ExperimenTation), and the technical details of the actual testbed. We then give the analysis in Section 4. Finally, our conclusions are in Section 5.

2 Requirements for Security ExperimenTation Environment

2.1 Risk Analysis and Risk Management

Depending on the kinds of research activities performed, security research can involve risks associated to the existing information infrastructure and ultimately to the public at large. From a confidentiality point of view, it is important to protect data from software vulnerabilities and malware collections from use for nefarious purposes. In addition, certain data collections that are used in security research, such as network traffic traces, could potentially contain private information. Similarly, details about network architecture or configuration of defensive mechanisms from real networks that are being emulated in a testbed facility should be protected from unauthorized access, as knowledge of these details could jeopardize their security. Most importantly, experiments involving live malware could "spill" into a real computing environment, affecting it from an integrity and availability point of view, for example, in the case of accidental releases of malware samples on previously uninfected machines. Alternatively, the effects of such actions could be indirect, such as those caused by researchers interacting with infected and criminal-controlled systems. In that case, this action could potentially trigger a premature and unnecessary arms race between the criminals and security researchers and security product vendors, by alerting cybercriminals that someone is "onto them", or on weaknesses in their tools and approaches.

2.2 Key Requirements Analysis

The construction of network experimental environment can be essentially conducted in four fashions: mathematical modeling, stochastic simulation, in laboratory emulations, or test-bed. The fidelity and scalability of the experimental environments built with different fashion are significantly different, and the laboratory emulation has often been the preferred method in security research, especially with regards to malware analysis.

Laboratory experimentation offers an interesting compromise. On the one hand, the controlled conditions in which they are ideally conducted provide 1) the ability to

validate previous experimental results obtained by others, i.e. repeatability, and 2) the ability to vary the parameters and characteristics of security solutions or threats being studied, i.e. experimental control. On the other hand, the use of the same or similar pieces of hardware (whether real or emulated) and the same software that is present in the real world, whether offensive or defensive, adds a level of a priori realism that is hard to attain just by modeling and simulation. However, in order to conduct security experiment, the emulation environment must obey the following criteria:

Scale. The number of elements (machines, subnets, processes, etc.) that are being emulated or recreated in the experiment should be large enough to approach the numbers in the real world or at the very least large enough so that statistically significant results can be obtained.

Fidelity. The static conditions describing the environmental setup should be as close as possible to those of typical equivalent environments in the real world. This includes network topology, server configurations, proportion of machines and equipment in various roles, and security mechanism.

Isolation. Security experiments in emulation environment are often involving live malware, such as worm and virus, which could destroy the security of the other network, such as management network or data-collection network. So the environment should isolate the live malware from the other network.

Sanitization. Even if the live malware could be isolated in experimental network, it can destroy subsequent experiment conducted in the environment. So, the experimental environment should sanitize machines and network nodes quickly and entirely.

3 Architecture of Security Experimental Environment

3.1 Main Idea

According to the requirement of security experiments, the LNESET are built on clustering, virtualization, and software routing technology. The architecture of LNESET is detailed in Fig. 1, and is basically composed of three layers; they are physical infrastructure layer, Meta resource layer, and Experiment layer.

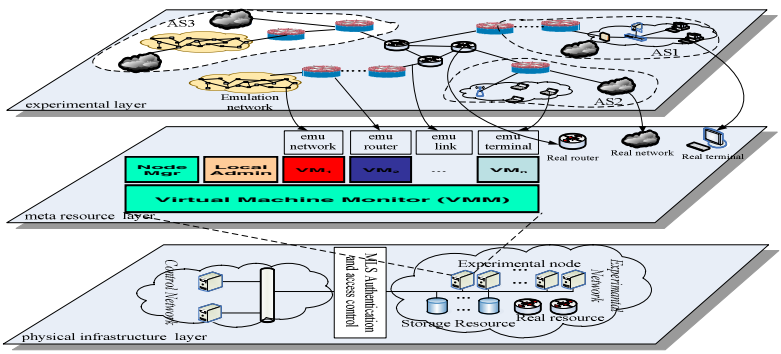


Fig. 1. The Architecture of Large-scale Network Emulation Environment for High-risk Security Experimentation

3.2 Physical Infrastructure Layer

At the lowest layer is the physical infrastructure layer, which is composed of compute nodes, switches, and other physical resources that is similar to the Emulab suite. The physical infrastructure of LNESET consists of two components: the control network and experimental network.

The control network consists of a management server and the data server, the former is responsible for the centralized management, configuration and monitoring of the hardware and software resources in the experimental environment. Besides, it provides a user interface to the researchers to configure, run, monitor the environment, furthermore collect and analyze the experiment results. The data server is used to store a variety of virtual image files and experimental data.

The experimental network is made up of various types of experiment nodes which are interconnected with the programmable network. The programmable network is stacked with some high-speed Ethernet switches which support VLAN partitioning. The experiment nodes include computer cluster, real router, real terminal and real network.

The control network and experimental network are isolated by MLS authentication and access control mechanism, which is very import for security experiment. It allows only a set of predefined information to flow between the control network and experimental network, which is needed to manage, configure, and sanitize the experimental network. Any other information will be denied to translate from experimental network to control network.

3.3 Meta Resource Layer

The meta resource layer is composed of a subset of resources allocated from the physical resource layer as needed by the experiment. After the requested resources have been successfully allocated by the control network, LNESET will bootstrap each compute cluster node by running a customized OS image that supports VMs.

The meta resources include not only real terminal, router and network connectivity between them, but also emulation resources which are constructed by virtualization on computer cluster node. These emulation resources include emulation network, router, terminal and link. LNESET realize the emulation network by extending NS2[8], realize emulation router by extending Quagga[9], and realize emulation Link by extending TC in Linux kernel[10]. In order to improve the scalability of computer cluster, LNESET use Xen[11] and OpenVZ[12], which could emulate one hundred nodes on one computer.

Meta resources are managed by LNESET. LNESET translates the experiment specification into meta resource requirements and instructs to allocate the resources for the experiment. After experiment, LNESET could retrieve resources by management server. In order to clear the live malware, the computer cluster node will reboot remotely on initial stage of each experiment, which will format the local disk. After that, LNESET will reload different emulation image from data server according experimental requirement, which will emulate network, router, terminal or link.

3.4 Experiment Layer

The experiment layer is created according to the network model of the experiment. Each compute cluster node serves as a basic scaling unit and implements the operation of emulation network, emulation terminal or emulation router that is mapped to it. A scaling unit for a network consists of a simulator instance and zero or more emulated hosts each running as a VM. An emulated host can also run directly on a physical compute node. This is designed for cases where the emulated hosts have stronger resource requirements and thus cannot be run as VMs. LNESET provides an emulation infrastructure to connect the emulated hosts running on virtual or physical machines to the corresponding simulator instances.

After the experiment starts, the experimenter can interact with the emulated hosts and the simulated network entities. To access the emulated hosts, the experimenter can directly log onto the corresponding computing nodes or VMs and execute commands (e.g., pinging any virtual hosts). To access the dynamic state of the simulated network, LNESET provides an online monitoring and control framework. This allows the experimenter to query and visualize the state of the network.

4 Analysis of Architecture

According to the previous architecture description in detail, LNESET can match the requirement of high-risk security experiment.

Scale. In LNESET, meta resources include emulation network, router, and terminal, which are realized by virtualization technology. Each computer cluster node can emulate network including one hundred node, such as router, terminal or server. So, LNESET could provide a large network environment for security experiment based on cluster, hence it has a better scalability.

Fidelity. In LNESET, except of real network, router and server, the emulation resources have high consistency with real resources. Based on virtualization technology, emulation resource could provide a realistic operating environment for testing real implementations of network protocol, applications and services. The routers, terminals and servers, and the connectivity between them could be configured according the experiment requirement. The network layer and system layer topology and network flow of environment could configure dynamically, so LNESET also provide a high level fidelity than test-bed based on simulation or analysis model.

Isolation. Because the MLS authentication and access control mechanism only allows very specific requests to proceed, the control network has very little interaction with the experimental network, so that it is not generally susceptible to attacks or malicious activities carried out by the security experiment such as worm or virus.

Sanitization. In LNESET, every resource will be initialized at the beginning of experiment. The experimental node include real node and cluster node will be sanitized entirely, because they reboot over network, and the boot files in server cannot rewrite. After reboot, the local disks of these experimental nodes will be formatted clearly. Every real router will be reset at the initial stage. So LNESET should sanitize all resources entirely.

5 Conclusion

Based on the in-depth analysis of the requirements of environment for high-risk experiment, in this paper we propose an emulation experimental environment named LNESET. We describe the architecture of LNESET in detail. LNESET supports high-level isolate between control network and experimental network, provides the necessary tools for the experimenters to allocate and reclaim the resources. The resources may consist of the compute nodes in the cluster and real machines or routers. Each compute node is treated as a scaling unit in LNESET; it can be configured as an emulation network with a set of VMs, a router or terminal for emulated hosts. The latter provides a realistic operating environment for testing real implementations of network applications and services. The VMs are connected with the simulator instances using a flexible emulation infrastructure. The analysis shows that LNESET is effective for high-risk security experimentation.

Acknowledgment. This research is supported by National Natural Science Foundation of China (Grant No. 61100223).

References

1. Maier, S., Herrscher, D., Rothermel, K.: Experiences with node virtualization for scalable network emulation. Computer Communications (30), 943–956 (2007)
2. PrimoGENI for hybrid network simulation and emulation experiments in GENI (2012)
3. Design and Evaluation of a Virtual Experimental Environment for Distributed Systems.
4. Simmonds, R., Unger, B.W.: Towards Scalable Network Emulation. Computer Communications 26(3), 264–277 (2003)
5. Network Emulator with Virtual Host and Packet Diversion (2012)
6. Design and Implementation of a Simulation Environment for Network Virtualization.
7. http://www.planet-lab.org
8. http://nanam.isi.edu/nanam/index.php/Main_Page
9. http://www.nongnu.org/quagga
10. http://www.kernel.org
11. http://xen.org
12. http://wiki.openvz.org/Main_Page
13. Ho, S.M., Lee, H.: A Thief among Us: The Use of Finite-State Machines to Dissect Insider Threat in Cloud Communications. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, JoWUA 3(1/2) (March 2012)
14. Hori, Y., Claycomb, W., Yim, K.: Guest Editorial: Frontiers in Insider Threats and Data Leakage Prevention. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, JoWUA 3(1/2) (March 2012)

Emulation on the Internet Prefix Hijacking Attack Impaction

Jinjing Zhao[1,2] and Yan Wen[1,2]

[1] Beijing Institute of System Engineering, Beijing, China
[2] National Key Laboratory of Science and Technology on Information System Security, Beijing, China
misszhaojinjing@hotmail.com,
celestialwy@gmail.com

Abstract. There have been many incidents of IP prefix hijacking by BGP protocol in the Internet. Attacks may hijack victim's address space to disrupt network services or perpetrate malicious activities such as spamming and DoS attacks without disclosing identity. The relation between network topology and prefix hijacking influence is presented for all sorts of hijacking events in different Internet layers. The impaction parameter is analyzed for typical prefix hijacking events in different layers. A large Internet emulation environment is constructed and the attack impaction of IP prefix hijacking events are evaluated. The results assert that the hierarchical nature of network influences the prefix hijacking greatly.

Keywords: IP prefix hijacking, Power law, BGP, Inter-domain routing system, Internet emulation environment.

1 Instruction

Prefix hijacking is also known as BGP hijacking, because to receive traffic destined to hijacked IP addresses, the attacker has to make those IP addresses known to other parts of the Internet by announcing them through BGP. Because there is no authentication mechanism used in BGP, a mis-behaving router can announce routes to any destination prefix on the Internet and even manipulate route attributes in the routing updates it sends to neighboring routers. Taking advantage of this weakness has become the fundamental mechanism for constructing prefix hijack attacks. They occur when an AS announces a route that it does not have, or when an AS originates a prefix that it does not own.

Previous efforts on prefix hijacking are presented from two aspects: hijack prevention and hijack detection. Generally speaking, prefix hijack prevention solutions are based on cryptographic authentications [4-8] where BGP routers sign and verify the origin AS and AS path of each prefix. While hijack detection mechanisms [9-15] are provided when a prefix hijack is going to happening which correction steps must follow. Because there is a lack of a general understanding on the impact of a successful

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 485–489, 2013.

prefix hijack, it is difficult to assess the overall damage once an attack occurs, and to provide guidance to network operators on how to prevent the damage.

In this paper, we conduct a systematic study on the impaction prefix hijacks launched at different position in the Internet hierarchy. The Internet is classified into three hierarchies—core layer, forwarding layer and marginal layer based on the commercial relations between autonomous systems (ASes). And a large Internet emulation environment is constructed which hybridizes the network simulation technology and packet-level simulation technology to achieve a preferable balance between fidelity and scalability. The experiment results show that the hierarchical nature of network influences the prefix hijacking greatly.

The remainder of this paper is organized as follows: The related works are discussed in section 2. The impaction analysis of the prefix hijacks attack is presented in section 3, in which IP prefix hijacks are classified on a comprehensive attack taxonomy relying on the Internet hierarchy model and BGP protocol policies. Section 4 builds an emulation environment to test the correctness of our conclusion and section 5 concludes the paper.

2 Related Work

Various prefix hijack events have been reported to NANOG [19] mailing list from time to time. IETF's rpsec (Routing Protocol Security Requirements) Working Group provides general threat information for routing protocols and in particular BGP security requirements [20]. Recent works [3,21] give a comprehensive overview on BGP security. The prefix hijacking is one of the key problems being noticed to BGP in these papers.

Previous works on prefix hijacking can be sorted into two categories: hijack prevention and hijack detection. The former one is trying to prevent the hijacking in the protocol mechanism level, and the latter one is trying to find and alert the hijacking event after it happening. The methods adopted can be categorized into two types: cryptography based and non-crypto based.

3 Analysis on Prefix Hijack Attack Impaction

3.1 Internet Hierarchy

In [18], we build a three-hierarchy model of the Internet and give an efficient arithmetic for it. The model is organized as follows:

a) The set of nodes who have no providers forms a clique (interconnection structure), which is the core layer.

b) If the nodes don't forward data for others, then it belongs to the marginal layer.

c) The node that belongs to neither the core layer nor the marginal layer belongs to the forwarding layer. And the forwarding layer has several sub-layers.

3.2 The Relation between Prefix Hijacking and the Internet Hierarchy

For the simpleness of the description, the ASes whose prefixes being hijacked are expresses with V, and the hijack attack ASes are denoted by A. Furthermore, we suppose each AS only has one provider. The multi-home mechanism is not considered in this paper.

To evaluate the influence if prefix hijacking events, two impaction parameters are introduced as follows:

Definition 1. Set of the affected nodes *Nc*: The set of nodes whose routing states might be changing because of the happening prefix hijacking event.

Definition 2. Affected path factor μ: The percentage of the paths might be changed because of the happening prefix hijacking event.

In paper [23], we classified the prefix hijacking events into nine types according to the different positions which the attackers and victims are located. The relation between prefix hijacking and the Internet hierarchy are concluded by the two impaction parameters.

From the analysis, these results can be drawn:

1) The hijacked AS in the core layer is not the most awful thing. On the contrary, if the AS in the marginal layer being hijacked, the number of the affected nodes is the largest among the three levels;

2) The hijacked AS in the forwarding layer can affect more paths than the core layer or the marginal layer;

3) If the hijacked ASes are in the same level, the hijacking AS in the forwarding layer can affect more nodes than the core layer or the marginal layer, and the higher attacker is in, the larger its influence will be;

4) The sub-prefix hijack can affect more ASes than the same prefix hijack, and the lager sub-prefix range is, the bigger affected path factor μ will be.

4 Evaluation Environment and Experiment

In order to verify the correctness of the conclusions in section 3, we build a prefix hijacking attack emulation environment, which is composed of three Juniper J2350 routers and four server computers. Each server can emulate 30 virtual routers.

For the authenticity of the test, the real BGP data is samples for the topology of inter-domain system. According to the sampling rules in [22], a network with 110 ASes is build, and the commercial relations are reserved. The network is also be classified into layers by the hierarchical algorithm in section 3.

Each prefix hijacking cases, we repeat the attach process three times, and calculate the average values of the affected nodes number *Nc* and path factor μ. The results are described in Table 1.

From the expariment results, we can see that if the AS in the marginal layer being hijacked, the number of the affected nodes is the largest among the three levels; the hijacked AS in the forwarding layer can affect more paths than the core layer or the marginal layer; and the hijacking AS in the forwarding layer can affect more nodes than the core layer or the marginal layer.

Table 1. Experiment Results

Case	*Nc*	*μ*
V∈C, A∈C	13	43
V∈C, A∈F	24	53
V∈C, A∈S	18	36
V∈F, A∈C	28	118
V∈F, A∈F	34	78
V∈F, A∈S	21	62
V∈S, A∈C	32	75
V∈S, A∈F	57	73
V∈S, A∈S	28	65

5 Conclusion

This paper conducts a systematic study on the impaction prefix hijacks launched at different position in the Internet hierarchy based on the work in paper [23]. A large Internet emulation environment is constructed which hybridizes the network simulation technology and packet-level simulation technology to achieve a preferable balance between fidelity and scalability. The experiment results show that the hierarchical nature of network influences the prefix hijacking greatly.

Acknowledgment. This research is supported by National Natural Science Foundation of China (Grant No. 61100223).

References

1. Lad, M., Oliveira, R., Zhang, B., Zhang, L.: Understanding Resiliency of Internet Topology Against Prefix Hijack Attacks. In: 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2007), pp. 368–377 (2007)
2. Nordstrom, O., Dovrolis, C.: Beware of BGP attacks. SIGCOMM Comput. Commun. Rev. 34(2) (2004)
3. Butler, K., McDaniel, P., Rexford, J.: A Survey of BGP Security Issues and Solutions. Proceedings of the IEEE 98(1) (January 2010)
4. Subramanian, L., Roth, V., Stoica, I., Shenker, S., Katz, R.H.: Listen and whisper: Security mechanisms for BGP. In: Proceedings of ACM NDSI 2004 (March 2004)
5. Ng, J.: Extensions to BGP to Support Secure Origin BGP (April 2004), `ftp://ftp-eng.cisco.com/sobgp/drafts/draft-ng-sobgpbgp-extensions-02.txt`
6. Kent, S., Lynn, C., Seo, K.: Secure border gateway protocol (S-BGP). IEEE JSAC Special Issue on Network Security (2000)
7. Zhao, S.S.M., Nicol, D.: Aggregated path authentication for efficient bgp security. In: 12th ACM Conference on Computer and Communications Security (CCS) (November 2005)
8. Smith, B.R., Murphy, S., Garcia-Luna-Aceves, J.J.: Securing the border gateway routing protocol. In: Global Internet 1996 (November 1996)

9. RIPE. Routing information service: myASn System, http://www.ris.ripe.net/myasn.html
10. Lad, M., Massey, D., Pei, D., Wu, Y., Zhang, B., Zhang, L.: PHAS: A prefix hijack alert system. In: 15th USENIX Security Symposium (2006)
11. Qiu, S., Monrose, F., Terzis, A., McDaniel, P.: Efficient techniques for detecting false origin advertisements in interdomain routing. In: Second workshop on Secure Network Protocols (NPSec) (2006)
12. Karlin, J., Forrest, S., Rexford, J.: Pretty good bgp: Protecting bgp by cautiously selecting routes. Technical Report TR-CS-2005-37, University of New Mexico (October 2005)
13. Xu, W., Rexford, J.: MIRO: multi-path interdomain routing. In: SIGCOMM 2006, pp. 171–182 (2006)
14. Hu, X., Mao, Z.M.: Accurate Real-time Identification of IP Prefix Hijacking. In: Proc. of IEEE Security and Privacy, Oakland (2007)
15. Zheng, C., Ji, L., Pei, D., Wang, J., Francis, P.: A Light-Weight Distributed Scheme for Detecting IP Prefix Hijacks in Realtime. In: Proc. of ACM SIGCOMM (August 2007)
16. Govindan, R., Reddy, A.: An Analysis of Internet Inter-Domain Topology and Route Stability. In: Proc. IEEE INFOCOM 1997 (March 1997)
17. Ge, Z., Figueiredo, D., Jaiwal, S., et al.: On the hierarchical structure of the logical Internet graph. In: Proceedings of SPIE ITCOM, USA (August 2001)
18. Zhu, P., Liu, X.: An efficient Algorithm on Internet Hierarchy Induction. High Technology Communication 14, 358–361 (2004)
19. The NANOG Mailing List, http://www.merit.edu/mail.archives/nanog/
20. Christian, B., Tauber, T.: BGP Security Requirements. IETF Draft: draft-ietf-rpsec-bgpsecrec-04 (March 2006)
21. Goldberg, S., Schapira, M., Hummon, P., Rexford, J.: How Secure are Secure Interdomain Routing Protocols? In: Proc. of ACM SIGCOMM, New Delhi, India, August 30–September 3 (2010)
22. http://www.ssfnet.org/Exchange/gallery/asgraph/src.tar.gz
23. Zhao, J.J., Yan, W., Xiang, L., et al.: The Relation on Prefix Hijacking and the Internet Hierarchy. In: The 6th International Conference on Innovative Mobile and Internet Services (IMIS 2012), Italy (July 2012)

Improved Clustering for Intrusion Detection by Principal Component Analysis with Effective Noise Reduction

Lu Zhao, Ho-Seok Kang, and Sung-Ryul Kim

Division of Internet and Multimedia Engineering, Konkuk University, Seoul, Korea
{ais.zhaolu,hsriver}@gmail.com, kimsr@konkuk.ac.kr

Abstract. PCA (Principal Component Analysis) is one of the most wildly used dimension reduction technique, which is often applied to identify patterns in complex data of high dimension [1]. In GA-KM [2], we have proposed GA-KM algorithm and have experimented using KDD-99 data set. The result showed GA-KM is efficient for intrusion detection. However, due to the hugeness of the data set, the experiment needs to take a long time to finish. To solve this deficiency, we combine PCA and GA-KM in this paper. The goal of PCA is to remove unimportant information like the noise in data sets which have high dimension, and retain the variation present in the original dataset as much as possible. The experimental results show that, compared to GA-KM [2], the proposed method is better in computational expense and time (through dimension reduction) and is also better in intrusion detection ratios (through noise reduction).

Keywords: Intrusion detection, Principle Component Analysis (PCA), effective noise reduction, GA-KM.

1 Introduction

With rapid growth of network-based services, network security is becoming more and more important than ever before. Therefore, intrusion detection system (IDS) [3] plays a vital role in network security. There are two main categories of intrusion detection techniques: signature-based detection and anomaly-based detection. Signature-based detection is also called misuse detection which is based on signatures for known attacks. Anomaly-based detection is different from signature-based detection, which is able to detect unknown attacks by learning the behavior of normal activity. In the training phase of this approach, IDS builds a profile which represents normal behavior. In the detection phase, the similarity of a new behavior with the profile is analyzed by IDS. If the new behavior is far from normal behavior of the profile, then this behavior will be labeled as an attack. We have proposed GA-KM algorithm [2], and have experimented with this algorithm using KDD--99 data set, the results show that it is efficient for anomaly-based intrusion detection [4]. However, when we experiment with this algorithm, it takes a long time to finish.

In this paper, to solve this deficiency, Principal Component Analysis is combined with GA-KM algorithm and is experimented on KDD-99 data set. PCA (Principal

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 490–495, 2013.

Component Analysis) can be used to reduce the dimension of high dimensional data and remove noise from it effectively. So we can use PCA algorithm to do some noise reduction on KDD-99 data. The experimental results show that the propose method reduces the training time and testing time greatly, and also is efficient for intrusion detection.

The rest of this paper is organized as follow: in section 2, we will introduce previous works. Section 3 describes proposed method. In section 4, we report our experimental results and evaluations. Finally, in section 5, we conclude this paper.

2 Related Work

2.1 GA-KM Algorithm

GA-KM algorithm [2] combines genetic algorithm into the traditional K-means clustering algorithm [5]. In GA-KM algorithm, each individual consists of a certain number of cluster centers. The value of each cluster center is called gene of individuals. And a fitness function is defined as follows:

$$f(x) = 1/(1 + J_c) \tag{1}$$

where J_c is the object function for K-means algorithm. We use this fitness function to evaluate each individual in a population. It means that if J_c is the small, the fitness value is the greater, and the clustering result is better.

In GA-KM, selection, crossover and mutation operators are a little different from original GA.

- Selection

We combine elitism selection and fitness proportionate selection [6] to increase the performance of GA-KM. Elitism selection first retains the best individual which has the best fitness value in current population, and then replaces the worst individual in new population with the best individual which is retained in current population.

- Crossover

To obtain better individuals and increase convergence speed of GA-KM, good individuals which have big fitness values cross over their genes among themselves and bad individuals which have small fitness values cross over their genes in arithmetic crossover method among themselves.

The selections of genes depend on two cluster centers' distance for two individuals, one cluster center for one individual. We first calculate the distance of all pairs of cluster centers each from two individuals, and then match up cluster centers based on their minimum distance and cross over the values of these two cluster centers.

In the formula (2), A'_1, B'_1 are genes of two good individuals that their cluster centers' distance is the minimum distance, A', B' are new genes of two individuals after crossover, and r is a random variable in the range [0,1].

$$A' = r * A'_1 + (1 - r) * B'_1$$
$$B' = (1 - r) * A'_1 + r * B'_1 \quad (2)$$

- Mutation

In this stage, the genes of bad individuals could have bigger chance of mutation than the genes of good individuals. So, in order to prevent early mature convergence and generate newer mutation of individuals, the genes of good individuals are modified based on original genes with a small value Min , and the genes of bad individuals are modified on original genes with a big value Max .

2.2 Principal Component Analysis

PCA (Principal Component Analysis Algorithm) known as dimension reduction technique is a useful method for identifying patterns in complex data of high dimensions, and expressing the data in such a way as to highlight their similarities and differences [7]. And this method can effectively identify data "main" elements and structure, remove noise and redundancy, and reveal the simple structure hidden in complex data.

3 Our Proposed Method

3.1 Data Description

For our experiments we use KDD-Cup 1999 dataset [8]. KDD-Cup 1999 data set contains a wide variety of intrusions simulated in a military network environment which is used for the Third International Knowledge Discovery and Data Mining Tools Competition. Each example in the data is a record of extracted features from a network connection gathered during the simulated intrusions. A connection is a sequence of TCP packets which data flows to and from a source IP address to a target IP address under some well-defined protocol. Each connection record consists of 41 fields which contain basic features about TCP connection as duration, protocol type, number of bytes transferred, domain specific features as number of file creation, number of failed login attempts, and whether root shell was obtained. Each connection is labeled as either normal, or an attack, with exactly one specific attack type.

3.2 Using PCA Algorithm

In this paper, the goal of PCA algorithm is to reduce unimportant information of data set which have high dimension and retain the variation present in the original dataset as much as possible. Using PCA algorithm to reduce the dimensionality of complex data can decrease computing time when we use huge data.

We use PCA algorithm to do effective noise reduction for data set S which are preprocessed by mapped. Here, data set S are represented as a $D \times N$ matrix, in the matrix, each column represents one data example of dataset; N is the number of data examples, D is the number of dimensionality for each connection record.

Firstly, calculate the mean value of all data examples:

$$y_{if} = \frac{1}{N} \sum_{i=1}^{N} s_{if} \tag{3}$$

Secondly, subtract mean vector for each data example:

$$y'_{if} = s_{if} - y_{if} \tag{4}$$

Thirdly, find covariance matrix $\sum$ of data set S, and then calculate all eigenvectors and eigenvalues of $\sum$;

Next, select number $d(d \le D)$ biggest eigenvectors based on eigenvalues to get robust representation of data, eigenvectors are represented as $u_1, u_2, \ldots, u_d$, and then find its matrix transpose U, represented as follows: $(d \times D \text{ matrix})$;

$$U = \begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_d^T \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1D} \\ u_{21} & u_{22} & \cdots & u_{2D} \\ \vdots & \vdots & \vdots & \vdots \\ u_{d1} & u_{d2} & \cdots & u_{dD} \end{pmatrix} \qquad y' = \begin{pmatrix} y'_{11} & y'_{21} & \cdots & y'_{N1} \\ y'_{12} & y'_{22} & \cdots & y'_{N2} \\ \vdots & \vdots & \vdots & \vdots \\ y'_{1D} & y'_{2D} & \cdots & y'_{ND} \end{pmatrix}$$

And then we can get the new data set $x = (Uy')^T$ preprocessed by mapping and PCA. In the matrix x, each row represents a data example, and the number of column is the same with the number of data examples.

4 Experiment Results and Evaluations

In this section, we detail the experiment results of the proposed algorithm and evaluate the performance of proposed method. We experimented with this proposed method using KDD Cup 99 dataset in our paper. For training and developing normal clusters, data file "kddcup.data_10_percent" which is downloaded from KDD Cup 1999 site is used for our experiments. The training data set consists of normal data selected from the file "kddcup.data_10_percent". And the testing data set consists of normal data and attack data which are selected from the file "corrected". The main goal of our experiment is to study how the performance of proposed method.

In our experiments, we use two kinds of data sets which are original KDD CUP 99 dataset and KDD CUP 99 dataset preprocessed by PCA algorithm with effective noise reduction. In [2], we have proved that GA-KM algorithm is efficient for anomaly-based intrusion detection. Our experimental results showed that the proposed method gives better representation of data set after removing noise data, and approximately 42.5% ~ 80.7% reduction in training time. Meanwhile, we compared the performance of improved clustering method with original GA-KM.

Table 1. Comparison with average distance

	PCA with GA-KM	GA-KM with scaling	GA-KM without scaling
Normal	1145.82495	3.143096903	1219.20607
Attack	1900.862749	3.796500426	1445.27273
Ratio1	1.66	1.21	1.19
Ratio2		1.37	

Table 2. Comparison with average distance with extreme cases removed

	PCA with GA-KM	GA-KM with scaling	GA-KM without scaling
Normal	1145.82495	3.143096903	1219.20607
Attack	1535.489161	3.225661856	1210.45372
Ratio1	1.34	1.02	0.99
Ratio2		1.12	

Table 3. Comparison with median distance

	PCA with GA-KM	GA-KM with scaling	GA-KM without scaling
Normal	1108.653748	2.764393351	1109.26483
Attack	1359.351694	3.197027967	1213.48725
Ratio1	1.23	1.15	1.09
Ratio2		1.07	

In the above three tables, scaling means zero-mean normalization method which is used to preprocess data set. And all distances means the minimum distance between cluster center and data. Normal represents the minimum distance between cluster center and normal data from testing data; Attack represents the minimum distance between cluster center and attack data; Ratio1 is the value which is equal to Attack divided by Normal; Ratio2 is the value which is equal to Ratio1 for PCA with GA-KM divided by Ratio1 for GA-KM with scaling. If the value Ratio1 is bigger, the difference between normal data and attack data is more evident, and detecting attack is easier. In other words, if the value Ratio2 is bigger, the performance of PCA with GA-KM is better than GA-KM with scaling.

In the experiments, we use three kinds of distances to compare the performance of proposed method and GA-KM. As shown in Table 1, we use average distance to compare PCA with GA-KM, GA-KM with scaling, GA-KM without scaling; the results show that the performance of PCA with GA-KM is 1.37 times better than GA-KM with scaling. As shown in Table 2, we remove the extreme distance and then calculate the average distance to compare these two methods, the value Ratio2 is 1.12, the result shows that the performance of proposed method is 1.12 times better than GA-KM with scaling. In Table 3, we use median distance; the result also shows that the performance of PCA with GA-KM is better than GA-KM with scaling. Also from the tables, we also can know that the performance of PCA with GA-KM, GA-KM with scaling are better than GA-KM without scaling.

From these results, we can know that performance of proposed method outperforms GA-KM. So we can know that after using PCA, the proposed method can reduce noise data effectively and retain the variation present in the original dataset as much as possible. Meanwhile, it also can retain a good performance after effective noise reduction. As a result the proposed method is also acceptable for intrusion detection.

5 Conclusion

Our research work is based on intrusion detection. We use KDD Cup 99 dataset to experiment and find some redundant and irrelevant data like noise data in KDD Cup 99. So to remove unimportant data effectively, we combine dimension reduction technique which is PCA algorithm with clustering method GA-KM. The goal of PCA is to reduce noise data effectively to get smaller dimensionality dataset from KDD Cup 99 dataset. Our research is to observe how Principal Component Analysis and GA-KM algorithm are used for intrusion detection. And when after the noise information of original dataset are reduced, how the performance is. The experimental results showed that the proposed method can give better and robust representation of data after effective noise reduction and have 42.5% ~ 80.7% time reduction in training, and also can retain a good performance. So, our proposed method is efficient and reliable for intrusion detection.

References

1. Smith, L.I.: A tutorial on Principal Components Analysis. New York (2002)
2. Zhao, L., Kang, H.-S., Kim, S.-R.: K-means Clustering by Genetic Algorithm for Anomaly-based Intrusion Detection. In: International Conference on Smart Media and Applications, p. 37. Korean Institute of Smart Media (2012)
3. Intrusion detection system, http://www.sans.org/reading_room/whitepapers/detection/understanding-intrusion detection-system_337
4. Denning, D.: An Intrusion Detection Model. In: Proceedings of the Seventh IEEE Symposium on Security and Privacy, vol. SE-13, pp. 222–232 (1987)
5. Kaufan, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, New York (1990)
6. Selection for genetic algorithm, http://en.wikipedia.org/wiki/Selection_(genetic_algorithm)
7. Lawrence, F.L., Sharma, S.K., Sisodia, M.S.: Network Intrusion detection by using Feature Reduction Technique. International Journal of Advanced Research in Computer Science and Electronics Engineering 1 (2012)
8. The third international knowledge discovery and data mining tools competition dataset\KDD99-Cup (1999), http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Unconditionally Secure Fully Connected Key Establishment Using Deployment Knowledge

Sarbari Mitra, Sourav Mukhopadhyay, and Ratna Dutta

Department of Mathematics
IIT Kharagpur, India
{sarbari,sourav,ratna}@maths.iitkgp.ernet.in

Abstract. We propose a key pre-distribution scheme to develop a well-connected network using deployment knowledge where the physical location of the nodes are pre-determined. Any node in the network can communicate with any other node by establishing a pairwise key when the nodes lie within each other's communication range. Our proposed scheme is unconditionally secure against adversarial attack in the sense that no matter how many nodes are compromised by the adversary, the rest of the network remains perfectly unaffected. On a more positive note, our design is scalable and provides full connectivity.

Keywords: sensor network, bivariate symmetric polynomial.

1 Introduction

Wireless Sensor Networks (WSN) are built up of resource-constrained, battery powered, small devices, known as sensors, which have capability of wireless communication over a restricted target field. Due to its immense application from home front to battle field, environment monitoring such as water quality control, landslide detection, air pollution monitoring etc., key distribution in sensor network has become an active area of research over the past decade. Usually sensor networks are meant to withstand harsh environments and thus secret communication is very essential. The secret keys are assigned to the nodes before their deployment in a Key Pre-distribution Scheme (KPS) to enable secure communication.

The bivariate symmetric polynomials were first used in key distribution by Blundo et al. [1]. The scheme is t-secure, i.e., the adversary cannot gain any information about the keys of the remaining uncompromised nodes if the number of compromised nodes does not exceed t. However, if more than t nodes are captured by the adversary, the security of the whole network is destroyed. Blundo's scheme is used as the basic building block in the key pre-distribution schemes proposed in [5, 6].

We present a deployment knowledge based KPS in a rectangular grid network by dividing the network in subgrids and applying Blundo's polynomial based KPS in each subgrid in such a way that nodes within communication range of each other can establish pairwise key. The induced network is fully connected

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 496–501, 2013.

– any two nodes, lying within communication range of each other, are able to communicate privately by establishing a secret pairwise key. The t-secure property of Blundo's scheme is utilized. A t-degree polynomial is assigned to at most $(t-1)$ nodes, where at least $(t+1)$ shares are required to determine the polynomial. This results in an unconditionally secure network, i.e., the network is completely resilient against node capture and this is independent of the number of nodes compromised. The nodes need to store at least $(t+1)\log q$ bits (where q is large prime) and a fraction of the total nodes needs to store at most $4(t+1)\log q$ bits. The storage requirement decreases with decreased radio frequency radius of the nodes. Comparison of the proposed scheme with existing schemes indicates that our network provides better connectivity, resilience and sustains scalability, with reasonable computation and communication overheads and slightly large storage for few nodes.

2 Our Scheme

Subgrid Formation: The target region is an $r \times c$ rectangular grid with r rows and c columns i.e., there are c cells in each row and r cells in each column of the grid. Each side of a cell is of unit length. A node is placed in each cell of the grid. Thus the network can accommodate at most rc nodes. Each of the $N(\leq rc)$ nodes are assigned a unique node identifier. All the nodes have equal communication range. Let ρ be the radius of communication range and d be the density of the nodes per unit length. Then $m = \rho d$ is the number of nodes lying within the communication radius of each node. We divide this network into a set of overlapping rectangular subgrids $SG_{i,j}$, for $i, j \geq 1$, of size $(2m+1) \times (2m+1)$ each. Each subgrid contains $(2m+1)^2$ cells and two adjacent subgrids overlap either in m rows or in m columns. By $N_{x,y}$ we denote the node at row x and column y in our rectangular grid. Deployment knowledge is used to get the idea about the location of the nodes after their deployment in the target field. We have designed the network to enable any pair of nodes lying in the radio frequency range of each other to be in at least one common subgrid. From the construction, t , according to our assumption. Let us assume that $R_i,\ 1 \leq i \leq r$ is the i^{th} row and $C_j,\ 1 \leq j \leq c$ is the j^{th} column of the rectangular grid. We refer a node to be *covered*, if it shares at least one common subgrid with each node within its communication range. Note that the nodes that lie at the intersection of the rows $R_i (1 \leq i \leq m)$ and columns $C_j\ (1 \leq i \leq m)$ are covered by subgrid $SG_{1,1}$. We consider sub-grid $SG_{1,2}$ and $SG_{2,1}$ overlap with $SG_{1,1}$ in m columns and m rows respectively, so that the nodes at the intersection of R_i and C_j, for $\{1 \leq i, j \leq 2m+1\} \setminus \{1 \leq i, j \leq m\}$, are made covered. Similarly, $SG_{2,2}$ intersects $SG_{1,2}$ and $SG_{2,1}$ in m rows and m columns respectively. This automatically covers all the nodes $N_{x,y}$ for $1 \leq x, y \leq 2m+2$. The overlapping of subgrids are repeated as described above to make all the nodes in the network covered.

Polynomial Share Distribution: Now, we apply Blundo's KPS in each subgrid. We choose randomly a bivariate symmetric polynomial $f_{ij}(x, y)$ of degree

$t > (2m+1)^2$ for subgrid $SG_{i,j}$, $i, j \geq 1$ and distribute univariate polynomial shares of the polynomial $f_{ij}(x, y)$ to each of the $(2m+1)^2$ nodes. Thus any node with identifier ID in subgrid $SG_{i,j}$ receives its polynomial share $P_{\mathsf{ID}}(y) = f_{ij}(\mathsf{ID}, y)$ and is able to establish pairwise keys with the remaining nodes in $SG_{i,j}$ following Blundo's scheme.

Now, let us discuss the scheme in detail for $m = 1$ in the following example.

Example: when $m = 1$

Lemma 21. The subgrid $SG_{i,j}$ consists of $(2m+1)^2 = 9$ nodes $N_{x,y}$, where $2i-1 \leq x \leq 2i+1$ and $2j-1 \leq y \leq 2j+1$.

Proof. From Figure 1, it follows that the result holds for $SG_{1,1}$. Without loss of generality, let us assume that the result is true for $i = i_1$ and $j = j_i$, i.e., the nine nodes of the subgrid SG_{i_1,j_1} are given by $N_{x,y}$, where $2i_1-1 \leq x \leq 2i_1+1$ and $2j_1-1 \leq y \leq 2j_1+1$.

Now we consider the subgrid SG_{i_1+1,j_1}. Each of the sub-grid are in the form of a 3×3 grid. From the construction it follows that the columns of SG_{i_1,j_1} and SG_{i_1+1,j_1} are identical, and they overlap in only one row (since $m = 1$), i.e., R_{2i_1+1}, which can also be written as $R_{2(i_1+1)-1}$. Therefore, SG_{i_1+1,j_1} consists of the nine nodes lying at the intersection of the rows $R_{2(i_1+1)-1}$, $R_{2(i_1+1)}$ and $R_{2(i_1+1)+1}$; and the columns C_{2j_1-1}, C_{2j_1} and C_{2j_1+1}. Thus the nodes of SG_{i_1+1,j_1} are given by $N_{x,y}$, where $2(i_1+1)-1 \leq x \leq 2(i_1+1)+1$ and $2j_1-1 \leq y \leq 2j_1+1$. Similarly, it can be shown that the rows of SG_{i_1,j_1} and SG_{i_1,j_1+1} are identical and they overlap in the column C_{2j_1+1}, which can also be represented as $C_{2(j_1+1)-1}$. Proceeding in the similar manner the nine nodes of the subgrid SG_{i_1,j_1+1} are $N_{x,y}$, where $2i_1-1 \leq x \leq 2i_1+1$ and $2(j_1+1)-1 \leq y \leq 2(j_1+1)+1$.

Thus the result holds for the subgrid SG_{i_1+1,j_1} and SG_{i_1,j_1+1}, whenever it is true for the subgrid SG_{i_1,j_1}. Also the result holds for $SG_{1,1}$. Hence, by the principle of mathematical induction, the result holds for subgrid $SG_{i,j}$, for all values of i, j. □

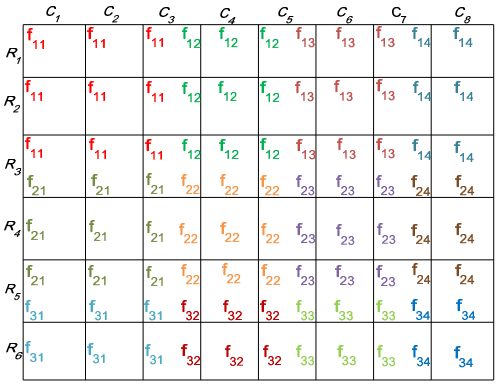


Fig. 1. Polynomial assignment to 3×3 overlapping sub-grid in a network, where $m = 1$

Lemma 22. Let univariate share of the bivariate symmetric polynomial f_{ij} be assigned to the node $N_{x,y}$.

(i) Let both x and y be even. Then $i = \frac{x}{2}, j = \frac{y}{2}$.

(ii) Let x be even and y be odd. Then $i = \frac{x}{2}$ and $j = \begin{cases} 1, & \text{if } y = 1; \\ \frac{y-1}{2}, \frac{y+1}{2}, & \text{otherwise .} \end{cases}$

(iii) Let x be odd and y be even. Then $j = \frac{y}{2}$ and $i = \begin{cases} 1, & \text{if } x = 1; \\ \frac{x-1}{2}, \frac{x+1}{2}, & \text{otherwise .} \end{cases}$

(iv) Let both x and y be odd. Then $\begin{cases} i = 1, j = 1, & \text{if } x = 1, y = 1; \\ i = 1, j = \frac{y-1}{2}, \frac{y+1}{2}, & \text{if } x = 1, y \neq 1; \\ i = \frac{x-1}{2}, \frac{x+1}{2}, j = 1, & \text{if } x \neq 1, y = 1; \\ i = \frac{x-1}{2}, \frac{x+1}{2}, j = \frac{y-1}{2}, \frac{y+1}{2}, & \text{otherwise .} \end{cases}$

Proof. From the construction of the scheme, it follows that univariate shares of the bivariate symmetric polynomial f_{ij} are distributed to each of the nine nodes of the subgrid $SG_{i,j}$. Thus our target is to find the coordinates of the subgrid $SG_{i,j}$ to which a node $N_{x,y}$ belong. Lemma 21 suggests that sub-grid $SG_{i,j}$ consists of the nodes $N_{x,y}$, for $2i - 1 \leq x \leq 2i + 1$ and $2j - 1 \leq y \leq 2j + 1$. Hence possible values of i are $\frac{x-1}{2}$, $\frac{x}{2}$ and $\frac{x+1}{2}$. Since i is an integer we must have $i = \frac{x}{2}$, when x is even and $i = \frac{x-1}{2}$ and $\frac{x+1}{2}$ when x odd. We further observe from Figure 1 that the first coordinate of all the subgrids and hence that of the corresponding bivariate polynomials assigned to the nodes lying in the first row is always 1. Similarly, possible values of j are $\frac{y-1}{2}$, $\frac{y}{2}$ and $\frac{y+1}{2}$, follows from Lemma 21. As j is also an integer we have $j = \frac{y}{2}$, when y is even and $j = \frac{y-1}{2}$ and $\frac{y+1}{2}$ when y is odd. We also observe from the Figure 1 that the second coordinate of all the subgrids and hence that of the corresponding bivariate polynomials assigned to the nodes lying in the first column is always 1, according to the construction of our design. Hence,

$$i = \begin{cases} 1, & \text{if } x = 1; \\ \frac{x}{2}, & \text{if } x \text{ is even;} \\ \frac{x-1}{2} \text{ and } \frac{x+1}{2}, & \text{otherwise,} \end{cases} \quad \text{and} \quad j = \begin{cases} 1, & \text{if } y = 1; \\ \frac{y}{2}, & \text{if } y \text{ is even;} \\ \frac{y-1}{2} \text{ and } \frac{y+1}{2}, & \text{otherwise .} \end{cases}$$

Hence, combining all the possible cases for the combination of the values of x and y we obtain the expression given in the statement of the Lemma. □

Theorem 23. We define the following variables for our $r \times c$ rectangular grid structure where K is the total number of symmetric bivariate polynomial required, M_1, M_3 and M_2 denote the total number of nodes containing only one, two or four polynomial shares respectively. We further identify the following cases as : *Case I* : – r and c both are odd; *Case II* : – r is odd and c is even; *Case III* : – r is even and c is odd and *Case IV* : – r and c both are even. Then

	K	M_1	M_2	M_3
Case I	$\frac{1}{4}(r-1)(c-1)$	$\frac{1}{4}(r+3)(c+3)$	$\frac{1}{4}(r-3)(c-3)$	$\frac{1}{2}(rc-9)$
Case II	$\frac{1}{4}(r-1)c$	$\frac{1}{4}(r+3)(c+2)$	$\frac{1}{4}(r-3)(c-2)$	$\frac{1}{2}(rc-6)$
Case III	$\frac{1}{4}r(c-1)$	$\frac{1}{4}(r+2)(c+3)$	$\frac{1}{4}(r-2)(c-3)$	$\frac{1}{2}(rc-6)$
Case IV	$\frac{1}{4}rc$	$\frac{1}{4}(r+2)(c+2)$	$\frac{1}{4}(r-2)(c-2)$	$\frac{1}{2}(rc-4)$

Proof. We provide the proofs in (a) and (b) for the expressions of K and M_1 respectively given in the table and leave the other two for page restrictions.

(a) According to the description of the scheme, each subgrid corresponds to a distinct bivariate polynomial, hence, the total number of polynomials required is equal to the total number of sub-grid present in the network. Let us assume that the sub-grid form a matrix consisting of r_1 rows and c_1 columns. Thus, we must have $K = r_1 c_1$.
This also follows from the construction that the subgrids are numbered in such a way that the coordinates of the k^{th} node of the subgrid $SG_{i,j}$ is less than or equal to the coordinates of the k^{th} node of the subgrid $SG_{i',j'}$, for $1 \leq k \leq 9$, whenever $1 \leq i \leq i' \leq r_1$ and $1 \leq j \leq j' \leq c_1$. According to the assumption, $N_{r,c} \in SG_{r_1,c_1}$. From Lemma 21 it follows that $2r_1 - 1 \leq r \leq 2r_1 + 1$ and $2c_1 - 1 \leq c \leq 2c_1 + 1$. Hence we must have $r_1 \geq \frac{r-1}{2}$ and $c_1 \geq \frac{c-1}{2}$. Since, r_1 and c_1 are integers (according to the assumption), we have

$$r_1 = \begin{cases} \frac{r-1}{2}, & \text{when } r \text{ is odd;} \\ \frac{r}{2}, & \text{when } r \text{ is even .} \end{cases} \text{ and } c_1 = \begin{cases} \frac{c-1}{2}, & \text{when } c \text{ is odd;} \\ \frac{c}{2}, & \text{when } c \text{ is even .} \end{cases}$$

Hence, considering the possible combinations from the above cases and substituting the values in the equation $K = r_1 c_1$, we obtain the expression given in the first column of the table given in the statement of the theorem.

(b) Let the node $N_{x,y}$ for $1 \leq x \leq r$, $1 \leq y \leq c$, stores exactly one univariate polynomial share. The possible values of x and y depends respectively on the number of rows r and number of columns c in the rectangular grid. Then it follows from the construction and can be verified from Figure 1 that

$$x \in \begin{cases} \{1, 2, \ldots, r\} \setminus \{2k+1 : 1 \leq k \leq \frac{r-3}{2}\}, & \text{when } r \text{ is odd;} \\ \{1, 2, \ldots, r\} \setminus \{2k+1 : 1 \leq k \leq \frac{r}{2} - 1\}, & \text{when } r \text{ is even .} \end{cases}$$

Hence, we get $\frac{r+3}{3}$ and $\frac{r+2}{3}$ possible cases for r being odd and even respectively. Similarly, we get $\frac{c+3}{3}$ cases when c is odd and $\frac{c+2}{3}$ cases when c is even. Hence, considering the possible combinations from the above cases and multiplying the corresponding values, we obtain the expression given in the second column of the table given in the statement of the theorem.

□

Resilience quantifies the robustness of the entwork against node capture. We consider the attack model as the random node capture, where the adversary captures nodes randomly, extracts the keys stored at them. Blundo's scheme has the t-secure property, as the adversary will not be able to gain any information if less than t nodes are compromised when univariate shares from a t-degree bivariate polynomial are assigned to the nodes. Here, we have assigned univariate shares of a t-degree bivariate polynomial where $t > (2m+1)^2$, to at most $(2m+1)^2$ nodes in a subgrid. Hence, even if upto $(2m+1)^2 - 2 = 4m^2 + 4m - 1$ nodes are captured by the adversary, the remaining two nodes will still be able to establish a pairwise key, which is still unknown to the adversary. This happens to all the pairwise independent bivariate polynomials. Hence, the network is unconditionally secure, i.e., no matter how many nodes are captured by the adversary, remaining network will remain unaffected.

Comparison: In Table 1, we provide the comparison of our scheme with the existing schemes proposed by Blundo et al. [1], Liu and Ning [6], Das and Sengupta [3] and Sridhar et al. [7]. Here, t denotes degree of the bivariate polynomial; q stands for the order of the underlying finite field $\mathbb{F}_q$; N is the total number of nodes in the network; s denotes the number of nodes compromised by the adversary and t in [3] is assumed to be sufficiently larger than $\sqrt{N}$, $c^{'}$ is a constant and $\mathcal{F}$ is the total number of polynomials in [6].

Table 1. Comparison with existing schemes

Schemes	Deployment Knowledge	Storage	Comm. Cost	Comp. Cost	Full Connectivity	Resilience	Scalable
[1]	No	$(t+1)\log q$	$O(\log N)$	$t+1$	Yes, 1-hop	t-secure	No
[6]	Yes	$c^{'}(t+2)\log q$	$c^{'}\log \lvert\mathcal{F}\rvert$	$t+1$	No	t-secure	No
[3]	No	$(t+2)\log q$	$O(\log N)$	$t+1$	Yes, 2-hop	secure	To some extent
[7]	No	$4(t+1)\log q$	$O(\log N)$	$O(t\log^2 N)$	No	depends on s	Yes
Ours	Yes	$4(t+1)\log q$	$O(\log N)$	$t+1$	Yes, 1-hop	secure	Yes

3 Conclusion

Utilizing the advantage of deployment knowledge and t-secure property of Blundo's polynomial based scheme, we design a network, which requires reasonable storage to establish a pairwise key between any two nodes within radio frequency range. The network is unconditionally secure under adversarial attack and can be scaled to a larger network without any disturbance to the existing nodes in the network.

References

1. Blundo, C., De Santis, A., Herzberg, A., Kutten, S., Vaccaro, U., Yung, M.: Perfectly-Secure Key Distribution for Dynamic Conferences. In: Brickell, E.F. (ed.) CRYPTO 1992. LNCS, vol. 740, pp. 471–486. Springer, Heidelberg (1993)
2. Chan, H., Perrig, A., Song, D.X.: Random Key Predistribution Schemes for Sensor Network. In: IEEE Symposium on Security and Privacy, pp. 197–213 (2003)
3. Das, A.K., Sengupta, I.: An Effective Group-Based Key Establishment Scheme for Large-Scale Wireless Sensor Networks using Bivariate Polynomials. In: COMSWARE 2008, pp. 9–16 (2008)
4. Das, A.K.: An Unconditionally Secure Key Management Scheme for Large-Scale Heterogeneous Wireless Sensor Networks. CoRR abs/1103.4678 (2011)
5. Li, G., He, J., Fu, W.Y.: A Hexagon-Based Key Predistribution Scheme in Sensor Networks. In: International Conference on Parallel Processing Workshops, ICPPW 2006, pp. 175–180 (2006)
6. Liu, D., Ning, P.: Improving Key Pre-Distribution with Deployment Knowledge in Static Sensor Networks. ACM Transactions on Sensor Networks 1(2), 204–239 (2005)
7. Sridhar, V., Raghavendar, V.: Key Predistribution Scheme for Grid Based Wireless Sensor Networks using Quadruplex Polynomial Shares per Node. Procedia Computer Science 5, 132–140 (2011)

An Improved Greedy Forwarding Routing Protocol for Cooperative VANETs*

Huaqing Wen and Kyung-Hyune Rhee**

Department of IT Convergence and Application Engineering,
Pukyong National University,
599-1, Daeyeon3-Dong, Nam-Gu, Busan 608-737, Republic of Korea
{wenhuaqing,khrhee}@pknu.ac.kr

Abstract. Most researches in the VANETs domain concentrate on the development of communication routing protocols. However, it is not effective to apply the existing routing protocols of MANETs to those of VANETs. In this paper, we propose a new greedy forward routing protocol which leverages real time traffic flow information to create a routing policy. Based on this routing policy, the proposed protocol alleviates the influence of high dynamic topology and decreases the average delivery delay on VANETs.

1 Introduction

Several routing protocols have been presented by many researchers for vehicular ad hoc networks (VANETs) [1, 7]. One of the best known position-based routing is GPSR [3]. GPSR works better in open space scenarios such as highways with evenly distributed nodes. In the condition of cities, GPSR suffers from many problems, because it does not consider some obstacles such as the movement of high-speed vehicles. In GPSR, when a node receives a Hello Message from its neighbors, it sets the Hello lifetime for each of its neighbors to prepare the next reception of this Hello Message. If it does not receive the Hello Message from its neighbor when the Hello lifetime expires, it decides that the neighbor has gone out of range. Due to the high mobility, a node may not receive updated location information from its neighbor since the neighbor has gone out of range [4]. Hence, when a node has data to forward, it may make a wrong decision which, in turn, leads to the packet loss. In order to provide a reasonable forwarding decision for nodes, we propose an improved greedy forwarding protocol based on the road traffic to alleviate the impact of vehicle mobility and enhance the stability of the routing link. We adopt both the fluid traffic model (FTM) [5] and intelligent driver model (IDM) [6]. FTM describes the speed as a monotonically decreasing function of the vehicle density and the speed of vehicle is calculated by means of the following equation:

* This research was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-013-D00121).

** Corresponding author.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 502–506, 2013.

$$S = max\{S_{min}, S_{max}(1 - k/k_{jam})\} \quad (1)$$

where S is the output speed, S_{min} and S_{max} are the minimum and maximum speed, respectively. k_{jam} is the vehicular density for which a traffic jam is detected. This last parameter is given by $k = n/l$, where n is the number of cars on the road and l is the length of the road segment itself. According to the equation (1), we can estimate the current flow speed S. Before we explain the protocol, in our system model, we assume that: (1) Each vehicle is equipped with Onboard Unit (OBU) and GPS device, which enable themselves to acquire their own positions and movement directions. (2) The source node already knows the current position of the destination node before transmission based on the location service. (3) All nodes are aware of the street-level information of the area where they are currently positioned.

2 Proposed Protocol

2.1 Neighbors Classification Phase

Each vehicle exchanges the information of the neighboring vehicles and updates the neighboring list table by Hello Messages. In order to improve the stability of the route, unlike the traditional neighbors selection that is based on the distance radius regardless of driving directions, in our approach, vehicles only with the same moving direction in both each road section and transmission range are set as one-hop neighbors or simply "Geographic Neighbors". In order to identify geographic neighbors for each vehicle, we insert vehicle's current speed and driving direction in Hello Message.

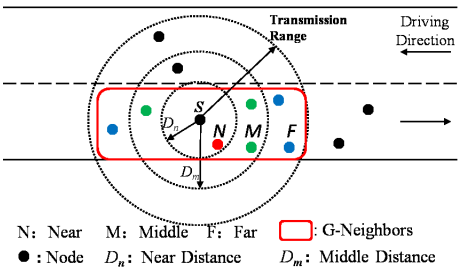


Fig. 1. Group Classification of Neighbors

Once each vehicle gets its geographic neighbor list, the node needs to calculate the current distance between geographic neighbors and itself. We define two preset parameters D_n and D_m to decide which group vehicles will belong to. If the distance D between vehicle and its neighbor is larger than the preset threshold D_m ($D_m \leq D$), this node will be classified into Far Group as shown in

Table 1. Example of Neighbor List Table

Neighbor List					
Node ID	Position x	Position y	Current Speed	Group Tag	Time Stamp
1	1234.56	4567.89	$12m/s$	N	3.0
2	1250.89	4564.12	$9m/s$	F	3.5

Figure 1. If $D_n \leq D < D_m$, neighbors in this area will be classified into Middle Group. If $D \leq D_n$, neighbors in this area will be classified into Near Group. After calculating this, vehicle attaches a group tag with each neighbor in the group list table. Every vehicle maintains a neighbor list table as shown in Table 1 and updates periodically.

2.2 Forwarding Policy and Packet Forwarding Phase

In equation (1), the number of node k is determined by information gathered from one-hop Geographic Neighbors. Because nodes are aware of the street-level information of the area where they are currently positioned, the length l of street can be measured. k_{jam} is available via a commercial navigation service, similar to the one currently provided by Garmin Traffic [8]. With the density information of the path, the traffic flow speed can be roughly estimated based on equation (1). We set a routing policy to determine which group will be chosen as the next hop candidate first. The priority level of each group is decided by the current speed as shown in Table 2.

Table 2. Routing Policy

Output Speed	Priority Level (from high to low)		
Slow	F	M	N
Medium	M	F	N
Fast	N	M	F

If the current flow speed S around the node is bigger than a preset S_{fast}, we will assign the Near Group as the first priority candidate and the Far Group as the last priority candidate. When the opposite happens, $S < S_{slow}$, the Far Group will be the first priority by reverse. For medium scenarios, $S_{slow} \leq S < S_{fast}$, we set the priority level from high to low as following: Middle Group, Far Group and Near Group. In the packet forwarding phase, the difference between the proposed greedy forwarding and GPSR is that we use a routing policy to choose the next hop so that the stability can be improved. When node S needs to send the first data packet of the event message, it will choose one which is the closest to the destination from the first priority group as the next hop forwarding packet. If the first priority group of the present node is empty, the node will check the second priority group. This process is repeated until all groups have been

checked. In the worst case, when node reaches a local maximum, the right hand rule to forward packets should be used.

Table 3. Analysis Parameters

Parameters	Measures
Road	Length: $1000m$
	Width: $6m$
Transmission Range	$250m$
Number of Nodes	30
Number of Nodes in Traffic Jam	300
Node Speed	$30km/h$ - $120km/h$
Hello Message Interval	$2.5s$
Packet Traffic	1 $packet/0.5s$
S_{slow}	$45km/h$
S_{fast}	$85km/h$

3 Analysis

Based on simulation result in [3], the routing overhead of GPSR is determined by Hello Message period. If we set long Hello Message exchange period, routing overhead will be decreased. However, with a long Hello Message interval packet, delivery ratio will be reduced because a next hop may go out of the communication range during a data packet transmission. The improved greedy forwarding routing protocol can compromise these two metrics when we set long exchange period. By using Matlab, we assume 30 nodes moving same direction are randomly distributed in one segment. Each node starts moving with a velocity ranges from $30km/h$ to $120km/h$ randomly and knows its current position as shown in Table 3. We compared 10 scenarios between GPSR and the proposed protocol performances. Figure 2 shows the data delivery ratio obtained for

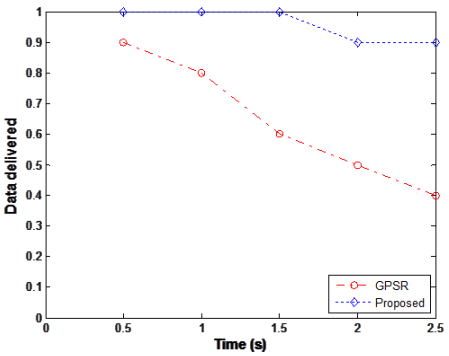


Fig. 2. Data Delivery Ratio

GPSR and the proposed protocol during one Hello Message interval. It is clearly shown that the proposed protocol guarantees a better delivery ratio compared to GPSR. Although the proposed protocol increases the hop from the source to destination, it avoids the case when a next hop goes out of the communication range during a data packet transmission and finally decreases the data loss and the resulting delay time, which can increase the routing performances.

4 Conclusion

In this paper, we proposed an improved greedy forwarding routing protocol. To design an efficient routing protocol for dynamic environment, the proposed protocol leverages real time vehicular traffic information to classify vehicle's neighbors and create a routing policy. Based on this routing policy, vehicles can select a reasonable node as the next hop. When density of traffic flow is low which indicates the speed in an area is high, and the topology changes frequently, a vehicle should forward packets to the neighbor which is near to avoid the case when a next hop goes out of the communication range during a data packet transmission. It enhances the stability of the routing link and decreases the average delay, which finally alleviates the influence of high dynamic of topology on VANETs.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Computer Networks, 393–422 (2002)
2. Boukerche, A., Oliveira, H., Nakamura, E., Loureiro, A.: Vehicular ad hoc networks: A new challenge for localization-based systems. Computer Communications 31(12), 2838–2849 (2008)
3. Karp, B., Kung, H.T.: GPSR: Greedy Perimeter Stateless Routing for Wireless Networks. In: Proc. of MOBICOM 2000, pp. 243–254 (August 2000)
4. Rao, S.-A., Boussedjra, M., Mouzna, J.: GPSR-L: Greedy Perimeter Stateless Routing with Lifetime for VANETS. In: Proc. of 8th International Conference on ITS, pp. 299–304 (October 2008)
5. Seskar, I., Maire, S., Holtzman, J., Wasserman, J.: Rate of Location Area Updates in Cellular Systems. In: Proc. of IEEE Vehicluar Technology Conference, vol. 2, pp. 694–697 (May 1992)
6. Treiber, M., Helbing, D.: Congested traffic states in empirical observations and microscopic simulation. In: Proc. of Statistical Mechanics, Physical Review E62, pp. 1805–1824 (2000)
7. Zhao, J., Cao, G.: VADD: Vehicle-Assisted Data Delivery in vehicular ad hoc networks. In: Proc. of the 25th INFOCOM, pp. 1–12 (2006)
8. Garmin Ltd. Garmin Traffic, `http://www8.garmin.com/traffic`

A Review of Security Attacks on the GSM Standard

Giuseppe Cattaneo[1], Giancarlo De Maio[1], Pompeo Faruolo[1], and Umberto Ferraro Petrillo[2,⋆]

[1] Dipartimento di Informatica *"R. M. Capocelli"*
Università di Salerno, I-84084, Fisciano (SA), Italy
{cattaneo,demaio,pomfar}@dia.unisa.it
[2] Dipartimento di Scienze Statistiche
Università di Roma *"La Sapienza"*, I-00185, Roma, Italy
umberto.ferraro@uniroma1.it

Abstract. The Global Systems for Mobile communications (GSM) is the most widespread mobile communication technology existing nowadays. Despite being a mature technology, its introduction dates back to the late eighties, it suffers from several security vulnerabilities, which have been targeted by many attacks aimed to break the underlying communication protocol. Most of these attacks focuses on the A5/1 algorithm used to protect over-the-air communication between the two parties of a phone call. This algorithm has been superseded by new and more secure algorithms. However, it is still in use in the GSM networks as a fallback option, thus still putting at risk the security of the GSM based conversations. The objective of this work is to review some of the most relevant results in this field and discuss their practical feasibility. To this end, we consider not only the contributions coming from the canonical scientific literature but also those that have been proposed in a more informal context, such as during hacker conferences.

Keywords: GSM, mobile security, security attacks, encryption.

1 Introduction

The GSM is the most widespread mobile communication technology, accounting for more than five billion subscriptions. Far from being just a personal communication technology, it has become the medium of choice for implementing and delivering a vast array of services ranging from mobile banking applications to electronic ticketing. This widespread use is also motivating the interest of researchers in evaluating the security mechanisms provided by GSM to protect user communication. In particular, the GSM protocols suffer from many weakness which allowed for the development of several attacks able to break confidentiality and privacy of subscribers. GSM carriers seem to have underestimated these threats, as witnessed by the several solutions for providing security to GSM-based communications (see [4,5,7,10]) proposed so far.

⋆ Corresponding author.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 507–512, 2013.

The objective of this paper is to review some of the most relevant security attacks to the GSM related technologies, including also those techniques that, although not being presented in a formal scientific context, have proved to be very effective in practice.

2 The GSM Standard

The GSM has been developed by the ETSI as a standard [1] to describe protocols for second generation digital cellular networks used by mobile phones. It offers several services based on voice transmission and data transmission. Three are the main elements of a GSM network. The first is the *Mobile Station*. It is made up of the Mobile Equipment (ME), the physical phone itself, and the Subscriber Identity Module (SIM). The SIM is a smart card that carries information specific to the subscriber together with the encryption keys (Ki and Kc). The second element is the *Core Network*. It carries out call switching and mobility management functions for mobile phones roaming on the network of base stations. It is made of several components. The third element is the *Base Station Subsystem*. It is responsible for handling traffic and signaling between a mobile station and the core network.

2.1 Security Features

The GSM standard defines several security mechanisms for protecting both the integrity of the network and the privacy of the subscribers. Whenever a ME tries to join a GSM network, it has to pass through an authentication procedure required to verify the identity of the subscriber using it. This denies the possibility for a subscriber to impersonate another one and guarantees that only authorized subscribers may access the network. When connected, the signaling and data channels over the radio path between a base station and the ME are protected by means of an encryption scheme. This ensures the confidentiality of the conversations. In the following we provide more details about these schemes and about the cryptographic machinery they use.

Authentication. The GSM network authenticates the identity of a subscriber using a challenge-response mechanism. Firstly, the Authentication Center (AuC), located within the core network, generates a 128-bit random number ($RAND$) and sends it to the ME. Then, the ME computes the 32-bit signed response ($SRES$) based on the encryption of $RAND$ with the authentication algorithm ($A3$) using the individual subscriber authentication key (Ki). The computation is entirely done within the SIM. This provides enhanced security, because the confidential subscriber information such as the individual subscriber authentication key (Ki) is never released from the SIM during the process. On the network, upon receiving the signed response ($SRES$) from the subscriber, the AuC compares its value of $SRES$ with the value received from the ME. If the two values match, the authentication is successful and the subscriber joins the

network. Notice that GSM authenticates the user to the network and not vice-versa. So, the security model offers confidentiality and authentication, but not the non-repudiation.

Data Confidentiality. The SIM contains the implementation of the key generation algorithm (A8) which is used to produce the 64-bit ciphering key (Kc) to be used to encrypt and decrypt the data between the ME and the base station. It is computed by applying the same random number ($RAND$) used in the authentication process to the ciphering key generating algorithm (A8) with the individual subscriber authentication key (Ki). Additional security is provided by the periodic change of the ciphering key. Similarly to the authentication process, the computation of the ciphering key (Kc) is done within the SIM.

Encrypted communications between the MS and the network is done using one of the A5 ciphering algorithms. Encrypted communication is initiated by a ciphering mode request command from the GSM network. Upon receipt of this command, the mobile station begins encryption and decryption of data using the selected ciphering algorithm and the ciphering key (Kc). The A5 algorithms are implemented in the hardware of the ME, as they have to encrypt and decrypt data on the fly.

The A5 Ciphering Algorithms. In the GSM protocol, the data is sent as sequence of frames, where each frame contains 228 bit. Each plaindata frame is XORed with a pseudorandom sequence generated by one of A5 stream cipher algorithms for ensuring over-the-air voice privacy.

The A5/1 algorithm was developed in late 1987 and is based on three *linear feedback shift registers* (LFSR). The keystream is built by running an algorithm, called *clock*, that produces 1 bit at each step. The output of the algorithm is the XOR of the leftmost bit of the three LFSR registers. Each register has a *clocking bit*. At each cycle, the clocking bits of the registers are given as input to a *function* that computes the *majority bit*. A register is *clocked* if the clocking bit agrees with the majority bit.

The A5/2 algorithm was introduced in 1989, it is a deliberate weakened version of the A5/1 that is almost identical to its counterpart except for an additional LFSR used to produce the three clocking bits. Since 2007, A5/2 is not implemented anymore in mobile phones for security reasons.

Finally, the A5/3 algorithm was developed in 1997 and is based on the MISTY cipher [12]. In the 2002 it was modified in order to obtain a faster and more hardware-friendly version, called KASUMI [1].

3 Attacks

There is a wide category of attacks against mobile communications that do not depend on network weaknesses. It includes mobile phones malware, identity theft by SIM cloning and so on. Some other attacks, such as phishing with SMS, may exploit human factors as well. A good review of such security issues can be found

in [6]. On the contrary, this work focuses on attacks that exploit vulnerabilities of GSM protocols.

Most of these attacks target the A5 family of ciphering algorithms. The exact formulation of these algorithms is still officially secret. However, the research community has been able to recover it through a mix of reverse engineering and cryptoanalysis. Namely, the general design of A5/1 was leaked in 1994 and the first cryptanalysis of A5/1 has been performed by Golic [9].

In this section we review some of the most interesting attacks proposed so far, distinguishing by passive and active attacks.

3.1 Passive Attacks

The first attack targeting the A5/1 algorithm has been proposed by Golic [9], which introduced an effective Time-Memory Trade-Off (TMTO) attack based on the birthday paradox. The basic idea of the TMTO is to pre-compute a large set of states A, and to consider the set of states B through which the algorithm progresses during the generation of output bits. Any intersection between A and B allows to identify an actual state of the algorithm. The proposed attack would be practicable only having 15 TB of pre-calculated data or 3 hours of known conversation [3].

Biryukov *et al.* presented two attacks based on a TMTO [3]. The first attack requires 2 minutes of known-conversation data and one second of processing time, while the second requires 2 seconds of plaintext data and several minutes of processing time. The amount of required storage is up to 290 GB. Unfortunately, its execution time grows exponentially with the decreasing of the input sequence. The attack exploits many weaknesses of A5/1, like the possibility of identifying states by prefixes of their output sequences, the ability to quickly retrieve the initial state of an intermediate frame and the possibility to extract the key from the initial state of any frame. The major drawback of this attack is the considerable amount of known-conversation data required.

A different strategy, based on a correlation attack, was introduced by Ekdahl *et al.* [8]. The main advantage is that whereas TMTO attacks have a complexity which is exponential with the shift register length, here the complexity is almost independent from it. This attack exploits the weakness that the key and the frame counter are initialized in a linear fashion, which enables to separate the session key from the frame number in binary linear expressions. This allows to decrypt a conversation in less than 5 minutes, provided that few minutes of plaintext conversation are available. Moreover, the time and space requirements for the tables precomputation are much smaller than in previous attacks.

All the attacks presented so far had very high computational cost and/or were based on unrealistic assumptions. Instead, the first practicable attack, implementable by means of open-source software and commodity hardware, has been made public by Nohl in 2009 [13]. This work showed that A5/1 is vulnerable to generic pre-computation attacks. In fact, for a cipher with small key (64 bit in the case of A5/1), it is possible to construct a *code book*. It can be exploited to perform a known-plaintext attack. For the case of A5/1, if an

adequate number of plaintext/ciphertext couples are known, it is possible to recover the encryption key. In the case of GSM, a number of predetermined control messages can be leveraged as known plaintexts [14]. Considering all the possible combinations, Nohl estimated that a code book for A5/1 would have been sized 128 PB and would have taken more than 100,000 years to be computed on a standard PC. In their talk, Nohl and Paget revisited techniques for computing the code book faster and for storing it compressed. They proposed a tweaked A5/1 engine optimized for parallelization. By using this technique a full code book for A5/1 can be computed in 3 months using commodity hardware. Some tweaks presented in subsequent talks [14,17] allowed to lower this boundary to 1 month on 4 ATI GPUs. Moreover, he proposed the use of a combined approach for data storage which makes use of distinguished point and rainbow tables [11], by means of which it is possible to reduce the size of the code book to just 2 TB.

Nohl estimates that the attack has a 99% success rate when data from a phone registered to the network can be collected, which maximizes the amount of known control frames. Otherwise, the success rate drops to 50%, since only a small number of frames with known plaintext is available. In a subsequent talk, Nohl and Munaut performed a demonstration on how it is possible to find phones and decrypt their calls [15].

3.2 Active Attacks

Differently from passive attacks, active attacks exploit some design weaknesses of the telecommunication infrastructure which make possible to introduce a false mobile tower controlled by the attacker. The major security hole exploited by the fake tower, also called IMSI Catcher, is that the GSM specification does not require authentication of the network to the mobile phone. The IMSI Catcher acts between the victim mobile phone(s) and the real towers provided by the service provider, and it is able to both control communication parameters, like encryption algorithms, and eavesdrop traffic. Such an attack falls into the category of Man-In-The-Middle (MITM) attacks.

Some MITM attacks against GSM have been introduced in [2]. They suppose that the victim is connected to a fake base station, which is able to intercept and forward the data sent by the phone to the network and vice versa. At this point, independently from the encryption algorithm chosen by the network, the attacker can request the victim to use a weak cipher like A5/2 (or even no encryption). Then, the attacker can employ cryptanalysis of A5/2 to retrieve the encryption key. It is worth noting that the key generation algorithm only depends on the RAND parameter specified by the network. As consequence, the encryption key used between the victim and the attacker is the same used between the attacker and the network, so that the attacker can decrypt all the traffic even if a secure encryption algorithm like A5/3 is requested by the network.

Paget and Nohl showed how it is possible to catch IMSI of a subscriber by means of an active attack [17]. Their attack makes use of a fake base station that could even be built from open source components.

In 2010 a practical attack to GSM has been presented by Paget [16] using open source components. It exploits the vulnerability that the mobile phone connects to the strongest base station signal. Since the base station has full control over communication protocols, the handset can be instructed in order to use no traffic encryption (A5/0). In this way, the attacker can intercept all the traffic in plaintext.

References

1. 3rd Generation Partnership Project (3GPP): Technical Specifications for GSM systems, http://www.3gpp.org/
2. Barkan, E., Biham, E., Keller, N.: Instant Ciphertext-Only Cryptanalysis of GSM Encrypted Communication. In: Boneh, D. (ed.) CRYPTO 2003. LNCS, vol. 2729, pp. 600–616. Springer, Heidelberg (2003)
3. Biryukov, A., Shamir, A., Wagner, D.: Real time cryptanalysis of A5/1 on a PC. Fast Software Encryption (2001)
4. Castiglione, A., Cattaneo, G., Maio, G., Petagna, F.: SECR3T: Secure End-to-End Communication over 3G Telecommunication Networks. In: IMIS 2011, pp. 520–526 (2011)
5. Castiglione, A., Cattaneo, G., Cembalo, M., De Santis, A., Faruolo, P., Petagna, F., Ferraro Petrillo, U.: Engineering a secure mobile messaging framework. Computers & Security 31(6), 771–781 (2012)
6. Castiglione, A., De Prisco, R., De Santis, A.: Do You Trust Your Phone? In: Di Noia, T., Buccafurri, F. (eds.) EC-Web 2009. LNCS, vol. 5692, pp. 50–61. Springer, Heidelberg (2009)
7. De Santis, A., Castiglione, A., Cattaneo, G., Cembalo, M., Petagna, F., Ferraro Petrillo, U.: An Extensible Framework for Efficient Secure SMS. IMIS 2010, pp. 843–850 (2010)
8. Ekdahl, P., Johansson, T.: Another attack on A5/1. Information Theory (2003)
9. Golić, J.D.: Cryptanalysis of Alleged A5 Stream Cipher. In: Fumy, W. (ed.) EUROCRYPT 1997. LNCS, vol. 1233, pp. 239–255. Springer, Heidelberg (1997)
10. GSMK: Cryptophone (2012), http://www.cryptophone.de/
11. Lee, G.W., Hong, J.: A comparison of perfect table cryptanalytic tradeoff algorithms. Cryptology ePrint Archive, Report 2012/540 (2012), http://eprint.iacr.org/
12. Matsui, M.: New Block Encryption Algorithm MISTY. In: Biham, E. (ed.) FSE 1997. LNCS, vol. 1267, pp. 54–68. Springer, Heidelberg (1997)
13. Nohl, K.: Subverting the security base of GSM. In: Hacking at Random (2009), https://har2009.org/program/events/187.en.html
14. Nohl, K.: Attacking phone privacy. In: BLACK HAT USA (2010), http://www.blackhat.com/html/bh-us-10/bh-us-10-archives.html
15. Nohl, K.: Wideband GSM sniffing. In: 27th Chaos Communication Congress (2010), http://events.ccc.de/congress/2010/Fahrplan/events/4208.en.html
16. Paget, C.: Practical Cellphone Spying. In: DEF CON 18 (2010), http://defcon.org/html/links/dc-archives/dc-18-archive.html
17. Paget, C., Nohl, K.: GSM: SRSLY? In: 26th Chaos Communication Congress (2009), http://events.ccc.de/congress/2009/Fahrplan/events/3654.en.html

An Extended Multi-secret Images Sharing Scheme Based on Boolean Operation

Huan Wang, Mingxing He, and Xiao Li

School of Mathematics and Computer Engineering,
Xihua University, 610039, Chengdu, China
{ideahuan18,hemingxing64}@gmail.com, lxgbxh@126.com

Abstract. An extended multi-secret images scheme based on Boolean operation is proposed, which is used to encrypt secret images with different dimensions to generate share images with the same dimension. The proposed scheme can deal with grayscale, color, and the mixed condition of grayscale and color images. Furthermore, an example is discussed and a tool is developed to verify the proposed scheme.

Keywords: Visual cryptography, Boolean operation, Image sharing, Multi-secret images.

1 Introduction

In traditional confidential communication systems, encryption methods are usually used to protect secret information. However, the main idea of the encryption methods is to protect the secret key [1]. The concept of visual cryptography is introduced by Naor and Shamir [2], which is used to protect the secret key.

Furthermore, there are a lot of works which are based on multiple-secret sharing schemes. Wang *et al.* [3] develop a probabilistic $(2, n)$ scheme for binary images and a deterministic (n, n) scheme for grayscale image. Shyu *et al.* [4] give a visual secret sharing scheme that encodes secrets into two circle shares such that none of any single share leaks the secrets. Chang *et al.* [5] report two spatial-domain image hiding schemes with the concept of secret sharing.

Moreover, many works are based on Boolean operation. Chen *et al.* [6] describe an efficient $(n + 1, n + 1)$ multi-secret image sharing scheme based on Boolean-based virtual secret sharing to keep the secret image confidential. Guo *et al.* [7] define multi-pixel encryption visual cryptography scheme, which encrypts a block of $t(1 \leq t)$ pixels at a time. Chen *et al.* [8] describe a secret sharing scheme to completely recover the secret image without the use of a complicated process using Boolean operation. Li et al. [9] give an improved aspect ratio invariant visual cryptography scheme without optional size expansion.

In addition, visual cryptography is used in some other fields. Wu *et al.* [10] propose a method to handle a secret image to n stego images with the $1/t$ size of the secret image. Yang *et al.* [11] design a scheme based on the trade-off between the usage of big and small blocks to address misalignment problem. Bose and

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 513–518, 2013.

Pathak *et al.* [12] find the best initial condition for iterating a chaotic map to generate a symbolic sequence corresponding to the source message.

These works are interesting and efficient but sometimes weakness, such as pixel expansion problems [2] and all the secret images should have the same dimension. However, generally, the secret images may have different dimension. Therefore, we propose an extended multi-secret images sharing scheme based on Boolean operation to encrypt multi-secret images with the different dimension. Moreover, the generated share images have the same dimension, then they do not reveal any information about the secret images include their dimension.

The rest of this paper is organized as follows. Section 2 gives the basic definitions. In section 3, an extended multi-secret images sharing scheme is proposed. An experimental is presented in Section 4. Section 5 concludes this paper.

2 Preliminaries

In this section, an extended-OR operation and an extended-OR operation chain between any two different dimensions images are defined. Let $x = 30$ and $y = 203$, then $x \oplus y = 00011110 \oplus 11001011 = 11010101 = 213$. Where, "$\oplus$" is bitwise exclusive-OR operation. Furthermore, The exclusive-OR operation between any two grayscale or color images with the same dimension is defined in [6].

Definition 1. *Let $A(a_{ij})$ and $B(b_{ij})$ be two images with* ***different dimensions*** *$m \times n$ and $h \times w$, respectively, where $m \times n \neq h \times w$, $0 \leq a_{ij} \leq 255, 0 \leq b_{ij} \leq 255$. The* ***extended-OR operation*** *between A and B is defined as follows.*

1) $A_{m\times n}\overline{\oplus'}B_{h\times w} = A_{m\times n} \oplus B'_{m\times n}$. Where, B' is a temporary matrix. If $m \times n \leq h \times w$, B' orderly takes $m \times n$ pixels from the head of B. Otherwise, B' circularly and orderly takes $m \times n$ pixels from the head of B.

2) $A_{m\times n}\overline{'\oplus}B_{h\times w} = A'_{h\times w} \oplus B_{h\times w}$. Where, A' is a temporary matrix. If $m \times n > h \times w$, A' orderly takes $h \times w$ pixels from the head of A. Otherwise, A' circularly and orderly takes $h \times w$ pixels from the head of A.

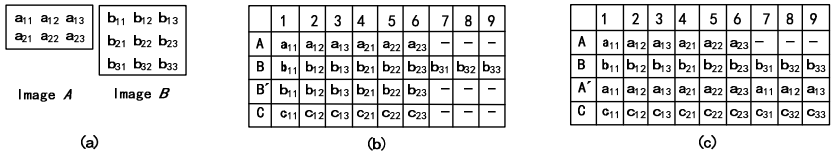


Fig. 1. An example for $\overline{\oplus'}$ and $\overline{'\oplus}$ operation

Example: Let $A_{2\times3}$ and $B_{3\times3}$ be two images, as shown in Fig.1 (a), the extended-OR operation between A and B are: $A_{2\times3}\overline{\oplus'}B_{3\times3} = A_{2\times3} \oplus B'_{2\times3}$, as shown in Fig.1 (b), and $A_{2\times3}\overline{'\oplus}B_{3\times3} = A'_{2\times3} \oplus B_{2\times3}$, as shown in Fig.1 (c).

Definition 2. *Let $A_1, A_2, \cdots, A_k$ be $k(k > 1)$ images with different dimensions. The* ***extended-OR operation chain*** *is defined as $\psi_{i=1}^{k} A_i = A_1 \overline{\oplus'} A_2 \overline{\oplus'} \cdots \overline{\oplus'} A_k$. Here, $A_1 \overline{\oplus'} A_2 \neq A_2 \overline{\oplus'} A_1$ unless A_1 and A_2 have the same dimension.*

3 The Sharing and Reconstruction of Multi-secret Images

In this section, n secret images with different dimensions can be encrypted to $n+1$ share images with the same dimension. $S_l, \cdots, S_m$ are denoted as $S_{[l,m]}$.

3.1 The Sharing Process

Sharing Algorithm: the sharing process is composed of following two parts.

Part1. For n secret images $G_{[0,n-1]}$, $n+1$ temporary images $S'_{[0,n]}$ with different dimensions are generated by following three steps.

(I) A random integer matrix is generated, which is the first temporary image S'_0 with the same dimension as G_1. Here, $\forall x \in S'_0$, $0 \leq x \leq 255$.

(II) According to S'_0 and the n secret images $G_{[0,n-1]}$, $n-1$ interim matrices $B_{[1,n-1]}$ are computed by $B_k = G_k \overline{\oplus'} S'_0$, $k = 1, 2, \cdots, n-1$.

(III) The other n temporary images $S'_{[1,n]}$ are computed by: a) $S'_1 = B_1$; b) $S'_k = B_k \overline{\oplus'} B_{k-1}$ if $k = 2, \cdots, n-1$; and c) $S'_n = G_0 \overline{\oplus'} B_{n-1}$.

Part2. $n+1$ share images $S_{[0,n]}$ with the same dimension can be generated by the $n+1$ temporary images $S'_{[0,n]}$ by the following steps.

(I) Extract the widths ($w_{[0,n-1]}$) and heights ($h_{[0,n-1]}$) of the n secret images $G_{[0,n-1]}$. Let $G^{wh}_{[0,n-1]}$ be n matrices with the same dimension 2×3, which are used to save the $w_{[0,n-1]}$ and $h_{[0,n-1]}$, respectively. We have:

$$G_i^{wh} = \begin{pmatrix} w_i^1 & w_i^2 & w_i^3 \\ h_i^1 & h_i^2 & h_i^3 \end{pmatrix}, where \begin{cases} w_i = w_i^1 \times w_i^2 \times w_i^3 \ \ , 1 \leq w_i^k \leq 255 \\ h_i = h_i^1 \times h_i^2 \times h_i^3 \ \ , 1 \leq h_i^k \leq 255 \end{cases}$$

Therefore, $G^{wh}_{[0,n-1]}$ can be considered as the new n secret images. Then, the new $n+1$ temporary images S_i^{wh} are generated from G_i^{wh} using Part1.

(II) According to $S'_{[0,n]}$ and S_i^{wh}, the $n+1$ share images $S_{[0,n]}$ can be computed as following steps.

(1) Let $M_w = max\{w_i\}$ and $M_h = max\{h_i\} + 1$.

(2) Generate $n+1$ empty images $S_{[0,n]}$ with dimension $M_w \times M_h$ and copy all the elements of $S'_{[0,n]}$ to $S_{[0,n]}$, respectively. The last lines of $S_{[0,n]}$ are empty.

(3) Copy all the elements of $S^{wh}_{[0,n]}$ to the last line of $S_{[0,n]}$, respectively.

(4) Fill in the rest of the $n+1$ images $S_{[0,n]}$ with the random numbers which are belong to 0 and 255.

Finally, the $n+1$ share images are generated with the same dimension $M_w \times M_h$. The proposed sharing scheme is shown in Fig.2.

Theorem 1. *Assume that n secret images $G_{[0,n-1]}$ with different dimensions are encrypted to $n+1$ share images $S_{[0,n]}$. All the share images cannot reveal any information independently.*

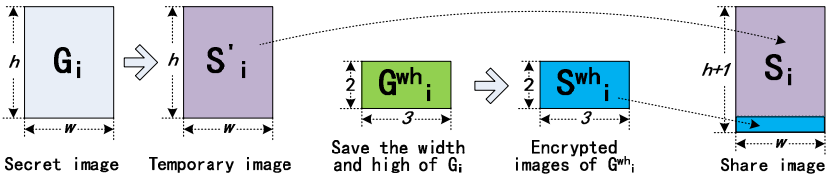


Fig. 2. Sharing process and the structure of share image

Proof: Since S_0 is a random matrix, then, obviously, $B_k = G_k{}'\overline{\oplus}S_0$ are still random matrixes. Furthermore, all S_k which are computed from B_k are also random matrices. Where, $k = 1, 2, \cdots, n-1$. Therefore, all the share images have the randomness, then they cannot leak any information independently.

3.2 The Reconstruction Process

Part1. The width and height of each secret image can be obtained from $S_{[0,n]}$.

(I) For the $n+1$ share images $S_{[0,n]}$, extract the $n+1$ temporary images $S^{wh}_{[0,n]}$ with the dimension 2×3 from the head of the last lines of $S_{[0,n]}$, respectively.

(II) The $n+1$ temporary images $S^{wh}_{[0,n]}$ can be decrypted using the following Part2 to obtain other $n+1$ temporary image $G^{wh}_{[0,n]}$ with the dimension 2×3.

(III) Let w_i^1, w_i^2, w_i^3 (h_i^1, h_i^2, h_i^3) be the first (second) line of G_i^{wh}, then $w_i = w_i^1 \times w_i^2 \times w_i^3$ $(h_i = h_i^1 \times h_i^2 \times h_i^3)$ is the width (high) of secret image G_i.

(IV) The $n+1$ temporary images $S'_{[0,n]}$ can be obtained from the $n+1$ share images $S_{[0,n]}$ according to the widths and highs in step III.

Part2. The n secret images $S_{[0,n]}$ can be obtained according to $S'_{[0,n]}$.

(I) The first secret image $G_0 = S_n\overline{\oplus'}B_{n-1} = S_n\overline{\oplus'}(S_{n-1}\overline{\oplus'}B_{n-2} = S_n\overline{\oplus'}(S_{n-1}\overline{\oplus'}(S_{n-2}\overline{\oplus'}B_{n-3})) = S_n(\overline{\oplus'}(S_{n-1}\overline{\oplus'}(S_{n-2}(\overline{\oplus'}, \cdots, (S_2\overline{\oplus'}S_1))\cdots)$.

(II) $n-1$ interim matrices B_k are generated by: $B_1 = S'_1$ and $B_k = S'_k\overline{\oplus'}B_{k-1}$, $k = 2, \cdots, n-1$.

(III) The other secret images are computed by $G_k = B_k\overline{\oplus'}S_0$, $1 \le k \le n-1$.

Theorem 2. *Assume that n secret images $G_{[0,n-1]}$ with different dimensions are encrypted to $n+1$ share images $S_{[0,n]}$, then the secret images $G_{[0,n-1]}$ can be correctly reconstructed using the $n+1$ share images $S_{[0,n]}$.*

Proof: If $k = 0$: We have $\Psi_{i=1}^{n} S_i = S_1\overline{\oplus'}S_2\overline{\oplus'}\cdots\overline{\oplus'}S_n = B_1\overline{\oplus'}(B_2\overline{\oplus'}B_1)\overline{\oplus'}\cdots\overline{\oplus'}(B_{n=1}\overline{\oplus'}B_{n-2})\overline{\oplus'}(G_0\overline{\oplus'}B_{n-1}) = G_0$. If $k \ge 1$: We have $\Psi_{i=0}^{k} S_i = S_0\overline{\oplus'}S_1\overline{\oplus'}\cdots\overline{\oplus'}S_k = S_0\overline{\oplus'}B_1\overline{\oplus'}(B_2\overline{\oplus'}B_1)\overline{\oplus'}\cdots\overline{\oplus'}(B_k\overline{\oplus'}B_{k-1}) = S_0\overline{\oplus'}B_k = G_k$.

3.3 Color Images and the Mixed Condition of Grayscale/Color Images

The difference between handling color and grayscale images is that each pixel of 24-bit color images can be divided into three pigments, i.e., red (r), green (g), and blue (b). We have $A\overline{\oplus'}B = [a_{i,j,k}\overline{\oplus'}b_{i,j,k}]$, where $k = r, g, b$.

For the mixed condition, each color image is divided into three (red, green, and blue) identical grayscale images. Let A be grayscale image and B be color image, we have $A\overline{\oplus'}B = [a_{i,j}\overline{\oplus'}b_{i,j,k}]$, where $k = r$ (red), g (green), b (blue).

4 Verification and Discussion

To verify the correctness of the proposed extended scheme, a tool is developed.

Example: There are five secret grayscale images G_0, G_1, G_2, G_3, G_4 with the dimensions 256×256, 360×477, 256×256, 196×210, 640×480, as shown in Fig.3(a). Here, $M_w = 640$ and $M_h = 480 + 1 = 481$. Then, the five secret images are encrypted and extended to six share images with the same dimension 640×481 using our tool, as shown in Fig.3(b). The reconstructed images are also decrypted using this tool, as shown in Fig.3(c). However, it is unsatisfactory that using these schemes developed in [3–12] to encrypt the five secret images since for any two secret images, some pixels in the bigger (dimension) secret image is out of the operation range for the smaller one and these pixels certainly cannot be encrypted. The comparison of these schemes is shown in Table 1.

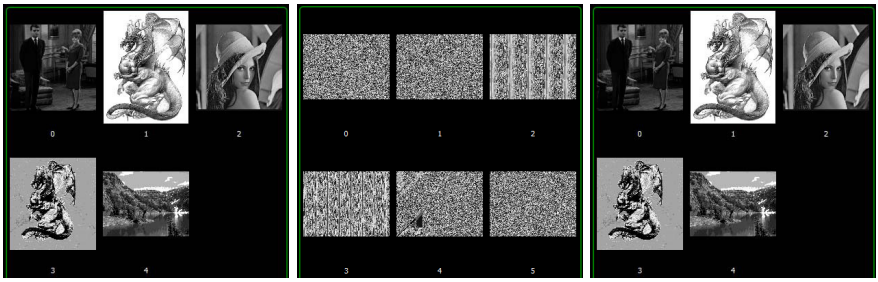


Fig. 3. An example with five secret images

Table 1. Comparison of these schemes

Schemes	Pixel expansion	Image distortion	Dimension restriction
In [3, 12]	No	Yes	Yes
In [4, 5]	Yes	Yes	Yes
In [6, 8]	No	No	Yes
In [7, 9]	Yes	No	Yes
In [10, 11]	No	Yes	Yes
This paper	No	No	No

5 Conclusions

An extended multi-secret image sharing scheme based on Boolean operation is proposed, which can share multi-secret images with different dimension. The grayscale and color images are appropriated in our scheme. Furthermore, this scheme can handle the mixed condition of grayscale and color images and the share images do not suffering pixel expansion. Moreover, the reconstructed secret images are the same dimension. In addition, all share images cannot leak any information about the secret images include the dimensions.

Acknowledgments. This work is supported by the National Nature Science Foundation of China (No. 60773035), the International Cooperation Project in Sichuan Province (No. 2009HH0009) and the fund of Key Disciplinary of Sichuan Province (No. SZD0802-09-1).

References

1. Shamir, A.: How to Share a Secret. Communications Assocating Computer 22(11), 612–613 (1979)
2. Naor, M., Shamir, A.: Visual Cryptography. In: De Santis, A. (ed.) EUROCRYPT 1994. LNCS, vol. 950, pp. 1–12. Springer, Heidelberg (1995)
3. Daoshun, W., Zhang, L., Ning, M., Xiaobo, L.: Two Secret Sharing Schemes Based on Boolean Operations. Pattern Recognition 40(10), 2776–2785 (2007)
4. Shyu, S.J.: Sharing Multiple Secrets in Visual Cryptography. Pattern Recognition 40(12), 3633–3651 (2007)
5. Chinchen, C., Junchou, C., Peiyu, L.: Sharing a Secret Two-tone Image in Two Gray-level Images. In: 11th International Conference on Parallel and Distributed Systems, pp. 300–304. IEEE Press, Tainan (2005)
6. Tzung, H.C., Chang, S.W.: Efficient Multi-secret Image Sharing Based on Boolean Operations. Signal Processing 6(12), 90–97 (2011)
7. Guo, T., Liu, F., Wu, C.: Multi-pixel Encryption Visual Cryptography. In: Wu, C.-K., Yung, M., Lin, D. (eds.) Inscrypt 2011. LNCS, vol. 7537, pp. 86–92. Springer, Heidelberg (2012)
8. Yihui, C., Peiyu, L.: Authentication Mechanism for Secret Sharing Using Boolean Operation. Electronic Science and Technology 10(3), 195–198 (2012)
9. Peng, L., Peijun, M., Dong, L.: Aspect Ratio Invariant Visual Cryptography Scheme with Optional Size Expansion. In: Eighth Intelligent Information Hiding and Multimedia Signal Processing, pp. 219–222. IEEE Press, Piraeus (2012)
10. Yus, W., Chihching, T., Jachen, L.: Sharing and Hiding Secret Images with Size Constraint. Pattern Recognition 37(7), 1377–1385 (2004)
11. Chingnung, Y., Anguo, P., Tseshih, C.: Misalignment Tolerant Visual Secret Sharing on Resolving Alignment Difficulty. Signal Processing 89(8), 1602–1624 (2009)
12. Bose, R., Pathak, S.: A Novel Compression and Encryption Scheme Using Variable Model Arithmetic Coding and Coupled Chaotic System. IEEE Transactions on Circuits and Systems-I: Regular Papers 53(4), 848–856 (2006)

Image Watermarking Using Psychovisual Threshold over the Edge

Nur Azman Abu, Ferda Ernawan, Nanna Suryana, and Shahrin Sahib

Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, Melaka 76100, Malaysia
{nura,nsuryana,shahrinsahib}@utem.edu.my, ferda1902@gmail.com

Abstract. Currently the digital multimedia data can easily be copied. Digital image watermarking is an alternative approach to authentication and copyright protection of digital image content. An alternative embedding watermark based on human eye properties can be used to effectively hide the watermark image. This paper introduces the embedding watermark scheme along the edge based on the concept of psychovisual threshold. This paper will investigate the sensitivity of minor changes in DCT coefficients against JPEG quantization tables. Based on the concept of psychovisual threshold, there are still deep holes in JPEG quantization values to embed a watermark. This paper locates and utilizes them to embed a watermark. The proposed scheme has been tested against various non-malicious attacks. The experiment results show the watermark is robust against JPEG image compression, noise attacks and low pass filtering.

Keywords: Image watermarking, JPEG image compression, edge detection.

1 Introduction

Currently, an efficient access internet makes it easy to duplicate digital image contents. In addition, current mobile devices view and transfer compressed images heavily [1]-[4]. Image watermarking is one of the popular techniques to manage and protect the copyright digital image content. Most of the image watermarking techniques exploits the characteristic of Human Visual System (HVS) in effectively embedding a robust watermark [5]-[7]. HVS is less sensitive to noise in highly textured area [8] and significantly changing region of an image. Human visual properties can be utilized in embedding process by insert more bits of watermark image for each block which has complex textures or edges on an image. The watermark with significant coefficients is more robust if it resides near round edges and texture areas of the image [9]. This paper proposes an embedding watermark scheme along the edge of the host image. This scheme enables the watermark to be more robust against non-malicious attacks.

The organization of this paper is given as follows. The next section will give a brief description on the concept of pshycovisual threshold for image watermarking. Section 3 presents an experimental design of the image watermarking. The

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 519–527, 2013.

experiment results of the propose watermark scheme are presented in Section 4. Section 5 concludes this paper.

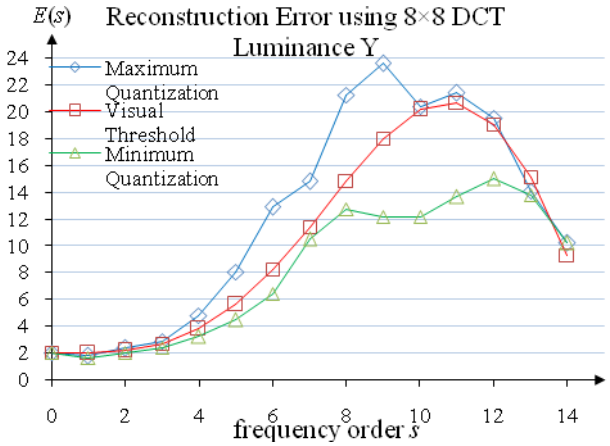


Fig. 1. Average reconstruction error of an increment on DCT coefficient on the luminance for 40 natural images

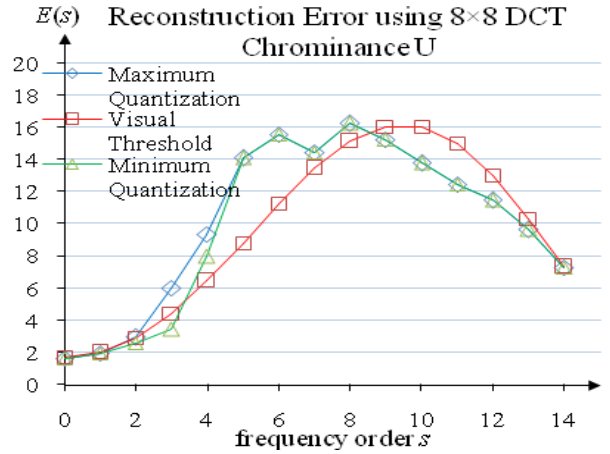


Fig. 2. Average reconstruction error of an increment on DCT coefficient on the chrominance for 40 natural color images

2 Psychovisual Threshold Based on Reconstruction Error

A psychovisual threshold has been generated from a quantitative experimental method. The image is divided into the 8×8 size blocks pixels and then transformed by 2-dimensional DCT. The resulting transformed coefficients are incremented one by

one on each frequency order. The DCT coefficients are incremented up to a maximum order of JPEG quantization table values. The effect of incrementing DCT coefficients are measured based on the image reconstruction error from the original image. The average reconstruction error from incrementing DCT coefficients for 40 natural images is shown in Fig. 1 and Fig. 2. The green line represents average image reconstruction error based on minimum quantization value of each order while the blue line represents average image reconstruction error based on maximum quantization table. The effect of incremented DCT coefficient as the frequency just noticeable difference of image reconstruction is investigated. The average image reconstruction error scores for each frequency order from order zero until order fourteen shall produce a transitional curve. In order to produce a psychovisual error threshold, the average reconstruction error is set by smoothing the curve line of image reconstruction errors as presented the red line. According to Fig. 1 and Fig. 2, there are gaps and loopholes in the popular JPEG quantization table. The gap is identified as the difference between minimum error reconstruction based on JPEG quantization table and an ideal error threshold reconstruction. We choose the gaps in the lower order because the watermark overthere is more robust against JPEG compression. The locations of loopholes in the popular JPEG quantization tables based on error threshold level are shown in Fig. 3.

16	14	13	15	19	28	37	55
14	13	15	19	28	37	55	64
13	15	19	28	37	55	64	83
15	19	28	37	55	64	83	103
19	28	37	55	64	83	103	117
28	37	55	64	83	103	117	117
37	55	64	83	103	117	117	111
55	64	83	103	117	117	111	90

18	18	23	34	45	61	71	92
18	23	34	45	61	71	92	92
23	34	45	61	71	92	92	104
34	45	61	71	92	92	104	115
45	61	71	92	92	104	115	119
61	71	92	92	104	115	119	112
71	92	92	104	115	119	112	106
92	92	104	115	119	112	106	100

Fig. 3. The locations if embedding watermark within 8×8 DCT coefficient for luminance (left) and chrominance (right) of new quantization tables Q_{VL} and Q_{VR} based on the psychvisual threshold

The watermark is expected to survive better in these deep holes against JPEG standard quantization tables Q_{CL} and Q_{CR} for luminance and chrominance respectively. These gaps can be computed as follows:

$$Q_{GL} = Q_{VL} - Q_{CL} \tag{1}$$

$$Q_{GR} = Q_{VR} - Q_{CR} \tag{2}$$

The location of embedding watermark image is in 8×8 DCT coefficient as depicted and blacken cell in Fig. 3. Each block is embedded watermark image randomly in blacken cells along the edge within the host image.

3 Experimental Design

An experimental sample uses a 'Lena' image 24-bit with size 512×512 pixel as a host image. The binary watermark image *W* "UTeM logo" has image size 32 × 32 pixels as shown in Fig. 4.

Fig. 4. Original watermark image consists of 32×32 pixels

The edge of a host image is computed using Canny edge detection to determine the edge location [11]. A sample Lena image is divided into blocks, each block consists of 8×8 image pixels. The host image is embedded a one-bit watermark for each edge image. The Random Numbers Generator (RNG) is an important computation in term of embedding process to generate the random numbers based on a secret key. The secret key is employed to encrypt and decrypt the watermark during watermark insertion and extraction. This paper implements mersenne twister method to generate random numbers for the embedding watermark scheme. —

3.1 Watermark Weight

The watermark that will be embedded into the host image is subjected to JPEG quantization table values. The quantization value that will be used as watermarks of embedding process is given as follows:

$$W_{QL} = \{18, 17, 16\} \text{ and } W_{QCR} = \{21, 26, 26\} \tag{3}$$

where the watermark weight depends upon the random numbers from a private key.

$$a = \begin{cases} 0.5 & \text{if } RNG(i,1) = 1 \text{ and } RNG(i,2) = 1 \\ 1 & \text{if } RNG(i,1) = 1 \text{ and } RNG(i,2) = 0 \\ 0.5 & \text{if } RNG(i,1) = 0 \text{ and } RNG(i,2) = 1 \\ 1 & \text{if } RNG(i,1) = 0 \text{ and } RNG(i,2) = 0 \end{cases} \tag{4}$$

The calculation of the watermark quantity is given as follows:

$$Q(i) = RNG(i) \cdot a \cdot W_Q \tag{5}$$

Consider a given 8×8 image block, if the watermark *W* is 1, the watermark image is multiplied by "+1" and added to the host image whereas the watermark *W* is 0, it is multiply by "−1" or subtracted from the host image. Each set occurs in the each pixel along the edge. This paper proposes an effective watermark embedding along the image edge randomly without degrading the visual image quality perceptually. If the watermarked image is disturbed along the boundary of an object in the image means

that the watermark will be degraded. Consequently, the host image lost its value whenever the visual aesthetic of the image was degraded.

3.2 Watermark Insertion

The most popular embedding watermark in frequency domain is in the most significant coefficient region [12]. It is a trade-off between robustness and imperceptibility. In this paper the watermark is embedded in the loopholes based on JPEG quantization tables. The embedding watermark is added along the edge of an object in the image. At the same time, a random process based on a key is used in the embedding a watermark or not along a given edge. The number of bits from watermark image to be embedded depends upon the numbers of edge image on each image block. The watermark weight will make sure that the watermark is perceptually invisible. This watermark procedure is as follows;

a. Take the host image block as an input (the size of an image block is 8×8 pixels).
b. Detect the edge image using Canny edge detection to the input image block.
c. Transform the image block by 2-D DCT if there are more than ten edges in image block.
d. Generate a unique random number based on a private key.
e. Determine the location for the watermark based on random numbers. The random numbers also used to generate the watermark weight which is one or half JPEG quantization value.
f. Embed the watermark into the loopholes when the block image has more than ten edges. Embedding value −1 or +1 when the watermark is 0 or 1 respectively.

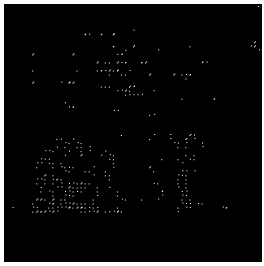


Fig. 5. An enhance absolute difference between the original image and the watermarked image

3.3 Watermark Extraction

The blind detection will be applied to extract the watermark image. This technique does not require an original host image for watermark detection. The watermark image can be obtained by using inner product approach. The watermark is extracted from the host image along the edge image [13]. The watermark image is dispersed randomly along the edge. The watermark is detected by computing correlation between the watermarked image and the watermark code. The difference between watermarked image and original image is enhanced and shown in Fig. 5. The extraction

of watermark involves a secret key to generate pseudo random numbers. This extraction is the inverse process of watermark embedding. The extracted watermark algorithms which includes the following steps:

a. Take the watermarked image block as an input (the size of image block is 8×8 pixels).
b. Detect of the edge image using Canny edge detection.
c. Transform the image block by 2-D DCT if there are more than ten edge image in image block.
d. Generate pseudo random numbers based on private key, these random numbers are used to find the location of watermarked image block.
e. Extracted watermark using inner product algorithm. In order to extract the extracted sequence of the $X^*\{x^*(i), (1 < i < N)\}$ where $x^*(i) \geq 1$ means that the watermark is 1 and $x^*(i) \leq 0$ means that the watermark is 0. The correlation coefficient in the below is used to decide if the watermark image exists in the image as follows:

$$\rho = X \cdot X^* \tag{6}$$

where $X \cdot X^*$ is the inner product of X and extracted sequence of the X^*. If the correlation coefficient between watermarked image X and extracted sequence X^* is larger than a threshold, we determine that watermark exists.

3.4 Image Watermarking Evaluation

In order to evaluate the concealing of the watermark, the standard Peak Signal to Noise Ratio (PSNR) is calculated to measure the quality of extracted watermark image. The comparison between the recovered watermark and the original watermark is quantitatively analysed by using Normalized Cross-Correlation (NC) [13], which is defined as follows:

$$NC = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} W(i,j).W'(i,j)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} W(i,j)^2 \sum_{i=1}^{M}\sum_{j=1}^{N} W'(i,j)^2}} \tag{7}$$

where the $W(i, j)$ is the original watermark image and $W'(i, j)$ is the recovered watermark image. $M \times N$ is the watermark image size and the value of NC is between 0 and 1. The higher value of NC means that the recovery watermark image is closer towards the original watermark image and more robust. In order to evaluate the performance of this watermarking scheme, the watermarked image undergoes various non-malicious attacks such as JPEG image compression, gaussian white noise, salt and pepper noise and gaussian low pass filter.

4 Experimental Results

This section examines the robustness of the image watermarking scheme against non-malicious attacks. Several experiments are tested to evaluate its performance of watermarked image and the results are shown in Table 1.

Table 1. Various Attacks for Watermarked Image

Various attacks	NC	PSNR
JPEG compression	0.8435	35.4610
Gaussian white noise 0.01	0.7183	20.2163
Salt & pepper noise 0.02	0.8003	22.1565
Gaussian low pass filter 3×3	0.9572	42.5371

The results of watermarked images and the corresponding extracted watermark after various non-malicious attacks are shown in Fig. 6 and Fig. 7.

Fig. 6. The results of watermarked image from (a) JPEG image compression, (b) Gaussian white noise 0.01, (c) Salt and pepper 0.02 and (d) 3×3 Gaussian low pass filter

(a) (b) (c) (d)

Fig. 7. The extracted watermark image after from (a) JPEG image compression, (b) Gaussian white noise 0.01, (c) Salt and pepper 0.02 and (d) 3×3 Gaussian low pass filter

Human visual system and its sensitivity are utilized in the design of this embedding watermarking scheme. The advantages of this watermarking scheme are the watermark image is perceptually invisible to the human eye and robust to non-malicious attacks. The destroying the image edge will remove the watermark. It also means that the quality of the host image will be damaged considerably.

The experimental results indicate that the watermark is resistance against non-malicious attacks. The recovered watermark changes slightly in comparison with original watermark and there is no significant change in visual perception between host image and watermarked image. Table I shows the comparison NC and PSNR of the proposed watermark against Gaussian low pass filter, JPEG compression, salt and pepper noise. The watermark scheme has a great resistance to salt and pepper noise. The results indicate that the proposed scheme is robust against the JPEG image compression. The extracted watermark image can be damaged but the visual perception of watermark still can be seen by human eye.

5 Conclusion

Digital image watermarking becomes an important technique for management and copyright protection. An effective secure embedding watermark based on human visual system properties is explored. The quantitative experiment of psychovisual error threshold level on natural images has been done. According to visual threshold model, there are gaps within the standard JPEG quantization tables. The embedding watermark image on loopholes based on JPEG quantization tables has been investigated. The watermark image is embedded along the edge based on psychovisual threshold visually invisible to human eye. The watermark is robust to non-malicious attacks. The experimental results show the extracted watermark image survives against JPEG image compression. In addition, the image watermarking scheme has strong resistance against added noise and low pass filtering.

Acknowledgment. The authors would like to express a very special thank to Ministry of Higher Education (MOHE), Malaysia for providing financial support for this research project via Fundamental Research Grant Scheme (FRGS/2012/FTMK/SG05/03/1/F00141).

References

1. Ernawan, F., Abu, N.A., Rahmalan, H.: Tchebichef Moment Transform on Image Dithering for Mobile Applications. In: Proceeding of the SPIE, vol. 8334, pp. 83340D-5. SPIE Press (2012)
2. Rahmalan, H., Ernawan, F., Abu, N.A.: Tchebichef Moment Transform for Colour Image Dithering. In: 4th International Conference on Intelligent and Advanced Systems (ICIAS 2012), pp. 866–871. IEEE Press (2012)
3. Ernawan, F., Noersasongko, E., Abu, N.A.: An Efficient 2×2 Tchebichef Moments for Mobile Image Compression. In: International Symposium on Intelligent Signal Processing and Communication System (ISPACS 2011), pp. 1–5. IEEE Press (2011)
4. Abu, N.A., Lang, W.S., Suryana, N., Mukundan, R.: An Efficient Compact Tchebichef Moment for Image Compression. In: 10th International Conference on Information Science, Signal Processing and their applications (ISSPA 2010), pp. 448–451. IEEE Press (2010)
5. Liu, K.C.: Human Visual System based Watermarking for Color Images. In: International Conference on Information Assurance and Security (IAS 2009), vol. 2, pp. 623–626. IEEE Press (2009)
6. Niu, Y., Kyan, M., Krishnan, S., Zhang, Q.: A Combined Just Noticeable Distortion Model-Guided Image Watermarking, Signal, Image and Video Processing, vol. 5(4), pp. 517–526. IEEE Press (2011)
7. Jayant, N.J., Johnston, J., Safranek, R.: Signal Compression Based on Models of the Human Perception. In: Proc. IEEE, vol. 81, pp. 1385–1422. IEEE Press (1993)
8. Yang, Y., Sun, X., Yang, H., Lie, C.T., Xiao, R.: A Contrast-Sensitive Reversible Visible Image Watermarking Technique. IEEE Transaction on Circuit and Systems for Video Technology 19(5), 656–667 (2009)
9. Bedi, S.S., Tomar, G.S., Verma, S.: Robust Watermarking of Image in the Transform Domain using Edge Detection. In: 11th International Conference on Computer Modelling and Simulation, pp. 233–238. IEEE Press (2009)
10. Abu, N.A., Lang, W.S., Sahib, S.: Image Projection over the Edge. In: International Conference on Computer and Network Technology (ICCNT 2010), pp. 344–348. IEEE Press (2010)
11. Canny, J.: A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8(6), 679–698 (1986)
12. Tseng, H.W., Hsieh, C.P.: A Robust Watermarking Scheme for Digital Images Using Self Reference. In: Advances in Communication Systems and Electrical Engineering, pp. 479–495. Springer, Heidelberg (2008)
13. You, X., Du, L., Cheung, Y., Chen, Q.: A Blind Watermarking Scheme Using New Nontensor Product Wavelet Filter Banks. IEEE Transaction on Image Processing 19(12), 3271–3284 (2010)

A Data Structure for Efficient Biometric Identification

Kensuke Baba and Serina Egawa

Kyushu University, Fukuoka, Japan
{baba.kensuke.060@m,s-egawa@soc.ait}kyushu-u.ac.jp

Abstract. This paper proposes an efficient algorithm for personal identification with biometric images. In identification based on image comparison, the number of comparisons is an important factor to estimate the total processing time in addition to the processing time of a single comparison. Maeda et al. proposed an identification algorithm that reduces the number of comparisons from the linear search algorithm, however the processing time of each comparison is proportional to the number of registered images. The algorithm in this paper is an improvement of the algorithm by Maeda et al. with constant-time image comparisons. This paper evaluates the algorithms in terms of the processing time and the accuracy with practical palmprint images, and proves that the novel algorithm can reduce the number of image comparisons from the linear search algorithm as the algorithm by Maeda et al. without loss of the accuracy.

1 Introduction

Personal authentication is an essential issue in many systems, especially, biometric authentication is an important technology to compensate some weaknesses of token- and knowledge-based authentication [6]. With the spread of computers and networks, the number of persons who use each application system is supposed to become huge. For authentication based on biometric information, there exist two possible procedures of matching, that is, verification and identification [6]. Identification searches for the target person, while verification confirms that the target person is a particular person. Identification requires a long processing time and it becomes more conspicuous in systems with a large number of users.

The aim of this paper is an acceleration of personal identification with biometric images. If biometric images are formalized as numerical vectors, the identification can be reduced to the problem of nearest neighbor search, that is, to search a set of vectors for the most similar vector to the input vector. Hence, an approach to the aim is an acceleration of nearest neighbor search by a suitable data structure [4] or an approximation [5]. In some practical systems, however, the process of image matching is implemented as a distinct module and treated as a black box whose input is a pair of two images and output is a similarity between the two images. In such a situation, identification should be conducted

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 528–533, 2013.

on the basis of image comparisons. In this paper, we focus on such comparison-based algorithms for identification with biometric images.

If we arrow some deterioration of accuracy, we can consider a method that searches for an image whose similarity with the input image is larger than a given threshold, instead of the most similar image. In this method, the processing time can be reduced by stopping the search when a similar image is found. Maeda et al. [9] proposed an identification algorithm (MSM) that reduces the number of image comparisons in the linear search. The main idea of the algorithm is that, for N images that were registered for authentication, a similarity between any pair of the registered images is calculated in advance, and then the order of comparisons with the input image is decided according to the $N \times N$ matrix of the similarities. They reported that the average number of comparisons is experimentally $O(\sqrt{N})$, while that in the linear search is $O(N)$. However, the process to choose the image for each comparison is comparing the N row vectors in the matrix, hence the processing time of a single image comparison is proportional to N. Therefore, the total processing time is estimated to be proportional to $N^{3/2}$ on the assumption that the number of comparisons is $O(\sqrt{N})$ [3].

In this paper, we propose an identification algorithm with biometric images as an improvement of MSM. The main idea is to prepare the order of image comparisons statically by a kd-tree [4] instead of the dynamic $O(N)$ computation from the $N \times N$ matrix. By the assumption of comparison-based identification, any algorithm requires $O(N)$ comparisons of images. The time complexity of the novel algorithm is no more than $O(N)$. Additionally, we examine the processing time and the accuracy of the three identification algorithms, the novel one, MSM, and the linear search, with the features extracted by Scale-Invariant Feature Transform (SIFT) [7,8] from practical palmprint images. We confirm that the novel algorithm reduces the number of image comparisons as MSM and does not worsen the accuracy.

2 Preliminaries

The problem of personal identification with biometric images is called *identification*. In identification, each image corresponds to a person. The input of identification consists of an image (called an *input image*) and a set of images (called a set of *templates*). The output is the name of the person judged to correspond to the input image or "null" if the person of the input image is judged to be not included in the persons of the templates.

In the rest of this paper, we consider comparison-based algorithms for identification. We suppose that an idea of similarity on biometric images is given and only the similarity of images can be obtained as leads for identification. Then, the *linear search algorithm* is, for a given threshold,

- Compare the input image with each template in the set successively in an order;
- If a template whose similarity with the input image is not less than the threshold is found, then output the person of the image and terminate;

– If the similarities with every templates are less than the threshold, output "null" and terminate.

For the accuracy of an identification algorithm, we consider "the rate that the person who corresponds to the output image is different from the person of the input image", and this rate is called the *error rate* (*ER*) of the algorithm.

3 The Algorithm

We propose an identification algorithm that reduces the number of image comparisons in the linear search algorithm by a tree for deciding the order of image comparisons.

Let N be the number of templates and t_i a template for $1 \leq i \leq N$. First, we make the $N \times N$ matrix whose (i, j)-element m_{ij} is the similarity between t_i and t_j for $1 \leq i, j \leq N$. Then, we construct a tree that decides the order of image comparisons as follows.

1. Choose ℓ templates for the initial comparisons;
2. Construct a kd-tree for the N ℓ-dimensional vectors obtained by 1;
3. Add an order of image comparison to each leaf of the tree in 2 on the basis of Euclidean distance on the ℓ-dimensional vectors.

Let $t_{r_1}, t_{r_2}, \ldots, t_{r_\ell}$ be the ℓ templates chosen in the process 1. In process 2, the kd-tree is constructed for the N ℓ-dimensional vectors

$$(m_{1r_1}, m_{1r_2}, \ldots, m_{1r_\ell}), (m_{2r_1}, m_{2r_2}, \ldots, m_{2r_\ell}), \ldots, (m_{Nr_1}, m_{Nr_2}, \ldots, m_{Nr_\ell}),$$

and each of the N leaves corresponds to a unique template. In process 3, an order of image comparisons is expressed as a list of the N templates, and the order for the leaf of t_i is decided as follows.

– The first template is t_i;
– The rest is the list of the templates except for t_i such that the jth template is nearer to t_i than $(j+1)$th template for $1 \leq j \leq N-2$,

where the distance between t_i and t_j is the distance between $(m_{ir_1}, m_{ir_2}, \ldots, m_{ir_3})$ and $(m_{jr_1}, m_{jr_2}, \ldots, m_{jr_3})$.

With the previous tree, identification is conducted as follows.

1. Compare the input image with the ℓ templates and obtain an ℓ-dimensional vector;
2. Search the nearest vector to the obtained vector in the constructed tree, and decide the order of comparisons;
3. Conduct the linear search algorithm with the decided order.

The process 1 requires ℓ comparisons of images. Generally, nearest neighbor search with a kd-tree in N ℓ-dimensional vectors needs $O(N^{1-1/\ell})$ time. In the process 3, since we have a fixed order of image comparisons, the processing time of a single comparison is constant. Therefore, the total processing time of the proposed algorithm is no more than $O(N)$. The practical number of image comparisons in the proposed algorithm is examined experimentally in the following section.

4 Experiments

The proposed algorithm in Section 3 was applied to practical palmprint images and evaluated in terms of the ER and the number of image comparisons.

4.1 Image Matching

We consider a matching of SIFT features for the comparison of palmprint images. This subsection defines the similarity on palmprint images for identification. In this paper, the region of interest on each palmprint was extracted as the circle that covers the maximal part on a palm as [3].

SIFT is one of the popular methods for image matching and object recognition and the detailed mechanism can be found in [7,8]. SIFT translates an image into a set of key points and each key point has a vector as its feature. Then, a comparison of two images is done by matching two sets of key points. There exist several possible procedures for the matching of key points. In this paper, the similarity on images (that is, sets of key points) is defined as follows. Let P and Q be two sets of key points and $v(p)$ the feature vector of a key point p. We consider q_p, p_q, and m such that

- For any $p \in P$, $q_p \in Q$ satisfies that $\|v(q_p) - v(p)\|$ is the smallest in Q.
- For any $q \in Q$, $p_q \in P$ satisfies that $\|v(p_q) - v(q)\|$ is the smallest in P.
- m is the number of the pairs of $p \in P$ and $q \in Q$ such that $q_p = q$ and $p_q = p$.

Then, the similarity of two images whose features are respectively P and Q is defined to be $m/\max\{|P|, |Q|\}$.

4.2 Results

The experiments were conducted on the PolyU Palmprint Database [2]. For the practical process of SIFT, the function "`SiftFeatureDetector`" in OpenCV [1] was used. The parameter "`threshold`" of the function was fixed at 0.01 and the other parameters were set to the default values according to the results of some preparatory experiments about the processing time and the error rate of matching.

In addition to the proposed algorithm, MSM [9] and the linear search algorithm were applied to a sample set of palmprint images. The set contains $1,200$ images that consists of 8 images times 150 persons. We separated the set into two sets of 4×150 images for templates and input images, and repeated each experiment with swapping the sets. The number of image comparisons in MSM depends on the choice of the pair of images for the first comparison, and that in the linear search depends on the order of templates in addition to the choice. Therefore, identification for the two algorithms were repeated for any combination of the initial pair (600×600 patterns), and moreover a cyclic order was chosen randomly and fixed for the repetition in the linear search. As for the proposed algorithm, the number of the repetition was 600.

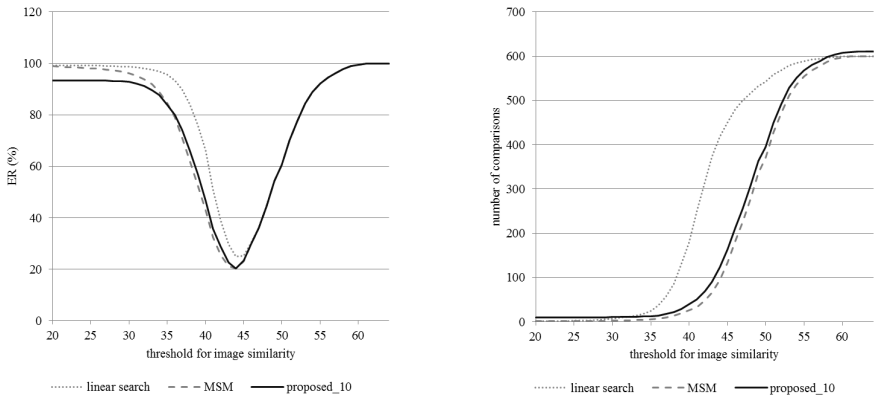


Fig. 1. The ERs and the numbers of image comparisons in the proposed algorithm, MSM, and the linear search algorithm

Table 1. The optimum ERs and the numbers of image comparisons at the point of the optimum ERs in the proposed algorithm, MSM, and the linear search algorithm. $*$ is the rate against the number in the linear search algorithm.

	ER	♯comparisons $(*)$
linear search	24.9	417.3 (1)
MSM	20.3	94.6 (0.23)
proposed	20.2	123.4 (0.30)

We fixed the number of initial comparisons $\ell = 10$ in the proposed algorithm and 10 initial templates were chosen randomly. Figure 1 shows the ERs and the number of image comparisons in the proposed algorithm with 10 initial comparisons, MSM, and the linear search algorithms at the different values of the threshold for the image similarity. The optimum ERs and the numbers of image comparisons at the optimum points are summarized in Table 1. The results report that the ER of the proposed algorithm is small compared to the linear search algorithm and almost same as MSM. The number of image comparisons in the proposed algorithm was drastically reduced from the linear search algorithm and is slightly larger than MSM.

5 Conclusion

This paper proposed an efficient algorithm for personal identification with biometric images as an improvement of the algorithm by Maeda et al. (MSM) [9]. We replaced the $O(N)$ process required for every image comparisons in MSM into an overhead of a single $O(N)$ process, where N is the number of the templates. The proposed algorithm, MSM, and the linear search algorithm were evaluated with practical palmprint images in terms of the processing time and

the accuracy. By the evaluation, we confirmed that the proposed algorithm can reduce the number of image comparisons from the linear search algorithm with no loss of the accuracy similarly as MSM.

In the experiment of the proposed algorithm, the number of initial comparisons was fixed to be 10, however the optimum number in the senses of the ER and the number of comparisons is not clear. To make clear the relation between the optimum number and the number of templates is one of our future work.

Acknowledgement. This work was partially supported by the Grant-in-Aid for Young Scientists (B) No. 22700149 of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) from 2010 to 2012 and CREST program of Japan Science and Technology Agency (JST) from 2008 to 2012.

References

1. OpenCV, `http://opencv.willowgarage.com/wiki/`
2. PolyU Palmprint Database, `http://www.comp.polyu.edu.hk/~{}biometrics/`
3. Egawa, S., Awad, A.I., Baba, K.: Evaluation of Acceleration Algorithm for Biometric Identification. In: Benlamri, R. (ed.) NDT 2012, Part II. CCIS, vol. 294, pp. 231–242. Springer, Heidelberg (2012)
4. Friedman, J.H., Bentley, J.L., Finkel, R.A.: An algorithm for finding best matches in logarithmic expected time. ACM Trans. Math. Softw. 3(3), 209–226 (1977)
5. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, STOC 1998, pp. 604–613. ACM, New York (1998)
6. Jain, A.K., Ross, A.A., Nandakumar, K.: Introduction to Biometrics. Springer, Heidelberg (2011)
7. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proc. IEEE International Conference on Computer Vision, pp. 1150–1157 (1999)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
9. Maeda, T., Matsushita, M., Sasakawa, K.: Identification algorithm using a matching score matrix. IEICE Transactions on Information and Systems E84-D(7), 819–824 (2001)

The PCA-Based Long Distance Face Recognition Using Multiple Distance Training Images for Intelligent Surveillance System

Hae-Min Moon[1] and Sung Bum Pan[2,*]

[1] Dept. of Information and Communication Engineering, Chosun Univ., Korea
bombilove@gmail.com
[2] Dept. of Control, Instrumentation, and Robot Engineering, Chosun Univ., Korea
sbpan@chosun.ac.kr

Abstract. In this paper, PCA-based long distance face recognition algorithm applicable to the environment of intelligent video surveillance system is proposed. While the existing face recognition algorithm uses the short distance images for training images, the proposed algorithm uses face images by distance extracted from 1m to 5m for training images. Face images by distance, which are used for training images and test images, are normalized through bilinear interpolation. The proposed algorithm has improved face recognition performance by 4.8% in short distance and 16.5% in long distance so it is applicable to the intelligent video surveillance system.

Keywords: intelligent surveillance system, image interpolation, long distance face recognition, principal component analysis(PCA).

1 Introduction

Recently, the video surveillance system has been developed to be intelligent, which includes finding criminals automatically or detecting fires by applying the techniques such as image analysis, computer vision or pattern recognition [1]. To satisfy the intelligent surveillance system, the long distance human identification technique applicable to the surveillance camera environment is needed [2]. The studies on long distance human identification using the face are still ongoing [3, 4]. Face recognition in the surveillance camera system should be operated in long distance(3m~5m) as well as short distance(1m~2m). When the existing face recognition algorithm applies to the surveillance camera system as it is, the recognition rate is reduced as the distance between the people and the camera increases. Recently, the technology that recognizes the long distance face using expensive camera which can obtain the high quality image from long distance is being studied [5, 6]. However, in case of face recognition, using expensive camera, it costs a lot to install and manage, making it difficult to use universally. Therefore, it is necessary to develop the long distance face recognition algorithm which can operate in the existing installed surveillance camera environment.

* Corresponding author.

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 534–539, 2013.

In this paper, PCA-based long distance face recognition algorithm applicable to the environment of surveillance camera is proposed. While single distance face images are used for training images for existing PCA-based face recognition, the proposed method uses face images by distance of 1m to 5m for the user training images. In addition, the size of face images extracted by distance are different for face images by distance of 1m to 5m, and thus the method is used to normalize the face images into the same size by using bilinear interpolation. As a result of experiment, the face recognition rate of existing algorithm was 86.6% in short distance and 48.4% in long distance, but the proposed face recognition algorithm showed 4.8% and 16.4% improved performance for 91.4% from short distance and 64.9% from long distance, respectively. The composition of this paper is as follows. In chapter 2, PCA, which is used in face recognition, and interpolation methods, which is used in normalization of face image size, are introduced. In chapter 3, the proposed long distance face recognition algorithm and experiment results are explained and chapter 4 concludes the paper.

2 Background

In this paper, PCA method which uses feature extraction method using basis vector are used [7]. To express two dimensional face images, face shape and texture information are vectorized. For face shape information, physiographic features like distance and ratio of face elements such as eye, nose and mouth are used. Texture information is expressed as brightness information itself in the face area by arraying the brightness value of two dimensional face images in order, features are extracted by expressing first-dimensional vector. The feature extraction process in face recognition is to find the base vector for linear transition. PCA technique is to find the eigenvector for covariance matrix as basis vector. PCA uses face images as a feature vector for face recognition by reflecting the face images to basis vector.

In case of long distance face recognition using PCA, the size of recognition images should be normalized fitting to the data size of training images. In general, since images taken from long distance are smaller than images taken from short distance, interpolation should be used to normalize the image size [8]. The nearest neighbor interpolation is the simplest method among interpolations and it refers to the pixel of nearest original images from the location that the output pixel is to be produced. Bilinear interpolation is a technique to produce the pixel to be interpolated using adjacent four pixels. The interpolated pixel is determined by the sum of four pixels multiplied by weighted value. At this time, weighted values are determined linearly and are inversely proportional to the distance from each of the adjacent pixels. Interpolation using higher-order polynomial equation defines the function of weighted value and is a method to calculate the pixel values by adding all values of neighboring pixel values of original images multiplied by weighted values. The representative method using higher-order polynomial equation includes cubic convolution interpolation. Bicubic convolution interpolation produces new interpolated pixels using 16 pixels of original images.

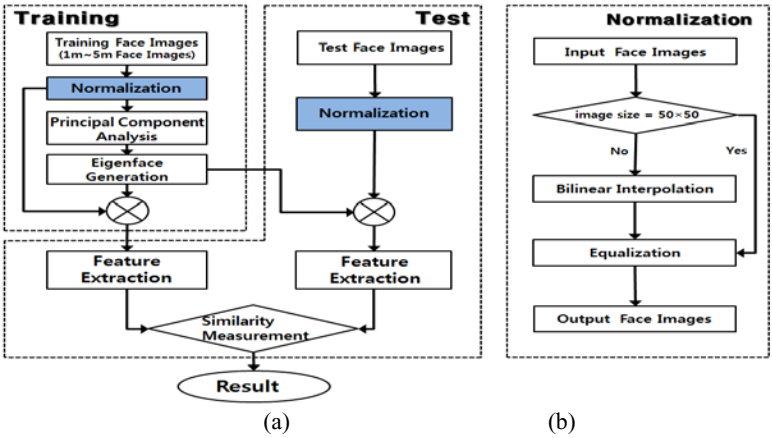


Fig. 1. Face recognition flowchart using PCA. (a) The flowchart of proposed long distance face recognition algorithm. (b) Flowchart of face image normalization.

3 Proposed Long Distance Face Recognition and Experimental Results

Fig. 1 is the flowchart of proposed PCA-based long distance face recognition. The overall flow of face recognition algorithm is same as existing PCA-based face recognition [7]. However, it has a difference in that proposed algorithm uses face images by distance of 1m to 5m as training images and adds the normalization process for face images by distance. Since PCA-based face recognition must use the difference of each face images and average face images, the size of images to be used in face recognition should be the same through the normalization. In this experiment, the size of face images was set as 50×50. 50×50 is the average size of face images extracted from 1m distance. The normalization process is shown in Fig. 1(b).

Face recognition algorithm is divided into training area and test area. Once the face images for training are entered, the size of input face images is judged. If the size of image is 50×50, the next step which is equalization will be conducted, but if the size is smaller than 50×50, the equalization will be conducted after enlarging the size to 50×50 through bilinear interpolation. All face images entered through this process will be normalized into 50×50 image size. Using normalized face images, the average face is produced in training images and by projecting each training image to the average face, the feature points unique to each face image are extracted. In test area like in training area, the input face images are normalized through interpolation and equalization. By comparing the feature points extracted from input face images to test area and feature points of face registered in training area, find the face images with the most similar values and classify.

Since the face recognition experimented in this paper needs face images by distance, the experiments in this paper used the ETRI face database [8] composed of face images by distance. ETRI face database obtained 500 face images (1m~5m: 100

images for each) per person from 10 people considering various lighting environment and distance change. The obtained face images were obtained through various lighting environment and 1m to 5m of distance change. To check whether the proposed method in this experiment is suitable for long distance face recognition situation, the experiment was conducted under the assumption that all of the face is extracted from input images by distance of 1m to 5m.

In this experiment, to select the appropriate interpolation for the normalization process of face images, 1m face images were used as training images and the face images by distance of 1m to 5m were used for test images. At this time, for normalization of face image size by distance of 1m to 5m, the nearest neighbor, bilinear, bicubic convolution and lanczos3 interpolations were used [10]. The original face image extracted from person 1 according to the distance change of 1m to 5m. The sizes of extracted face images are 50×50, 30×30, 20×20, 16×16 and 12×12 from 1m to 5m, respectively. The face images extracted by distance are normalized by four kinds of interpolation.

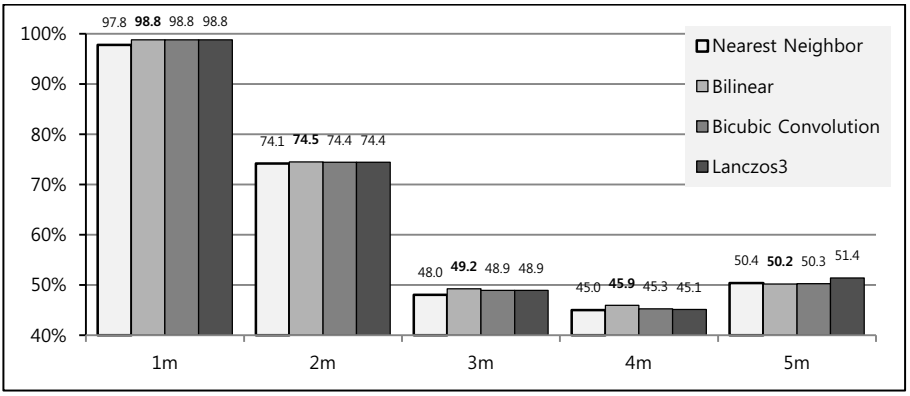


Fig. 2. Face recognition rate by distance according to interpolation

For training images per one person, 20 images of 1m face image were used and 80 images of face image by distance of 1m to 5m were used for test images. Fig. 2 shows the change of face recognition rate by distance according to interpolation in the same training and test condition. As a result of experiment, for short distance, when bilinear interpolation was used, the short distance face (1m to 2m) had the best recognition performance which is 86.6%. When the nearest neighbor interpolation, bilinear, bicubic convolution and lanczos3 interpolations were used, the long distance (3m to 5m) showed similar performances which were 47.8%, 48.4%, 48.2% and 48.5%, respectively. Therefore, the bilinear interpolation was used in PCA-based face recognition considering the complexity of computation, execution time and performance.

The existing face recognition algorithm has only used the single distance face images for training images, but the proposed algorithm improves the face recognition rate using the face images by distance for training images. Table 1 is the experimental

condition to compare the recognition rate when using single distance face images and face images by distance for training images. In CASE 1, only images taken at 1m were made up for training images and the number of training images per person used as 20 images. In CASE 2, the number of entire training images per person was 20 images which is same as CASE 1, but instead of using 1m face images, total of 20 images with 4 images each by distance of 1m to 5m were used. Fig. 3 shows the change of face recognition rate according to construction of training images. The experiment was conducted under the same condition that the number of training images per person was 20 images. When using single distance for training images, the performance was shown for 86.6% in short distance and for 48.4% in long distance. However, when using face images by distance of 1m to 5m, the short distance had better performance for 91.4% than when using single distance for training images which is 64.9%.

Table 1. Face recognition experiment according training images

CASE	Training Condition		
1	Training image per person	-1m	: 20 images
	Test image per person	-1m~5m	: 80 images each
2	Training image per person	-1m~5m	: 4 images each
	Test image per person	-1m~5m	: 80 images each

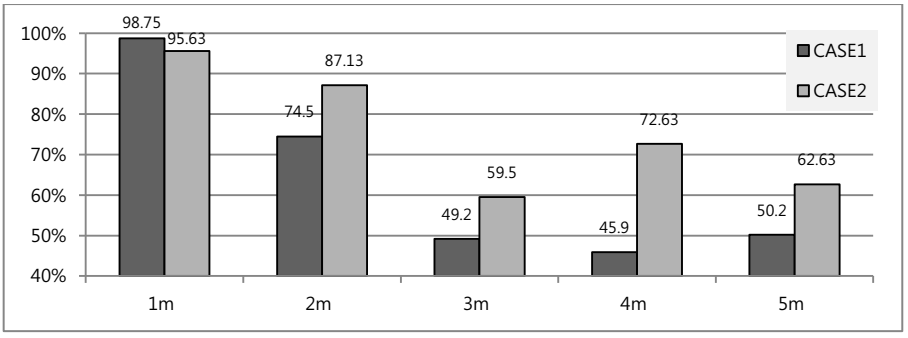


Fig. 3. Face recognition rate by distance according construction of training images

4 Conclusions

As various incidents are frequently occurred recently, the interest in long distance human identification technology is also increasing with the development of intelligent video surveillance camera. The PCA-based face recognition which has used the existing single distance face images as training images has disadvantage of lowering the recognition rate as the distance between surveillance camera and the user increases. In this paper, PCA-based long distance face recognition algorithm that is applicable to the environment of surveillance camera is proposed. The proposed face recognition

algorithm uses face images by distance of 1m to 5m for training images and the bilinear interpolation is used for normalization of input face images by various distances. As a result of experiment, the proposed face recognition algorithm had improved face recognition rate by 4.8% in short distance and by 16.5% in long distance.

In the future, to improve the inconvenience of registering face images by distance of 1m to 5m by the user, the study to produce the various face images by distance automatically will be conducted using single distance images.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0023147).

References

1. Aramvith, S., Pumrin, S., Chalidabhongse, T., Siddhichai, S.: Video Processing and Analysis for Surveillance Applications. In: Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems, Kanazawa, Japan, pp. 607–610 (2009)
2. Moon, H.M., Pan, S.B.: A New Human Identification Method for Intelligent Video Surveillance System. In: Proceedings of 19th International Conference on Computer Communication and Networks, Zurich, Switzerland, pp. 1–6 (2010)
3. Yi, Y., Abidi, B., Kalka, N.D., Schmid, N., Abidi, M.: High Magnification and Long Distance Face Recognition: Database Acquisition, Evaluation, and Enhancement. In: Proceeding of 2006 Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference, Baltimore, USA, pp. 1–6 (2006)
4. Tsai, H.C., Wang, W.C., Wang, J.C., Wang, J.F.: Long Distance Person Identification using Height Measurement and Face Recognition. In: Proceeding of IEEE Region 10 Conference TENCON 2009, Singapore, pp. 1–4 (2009)
5. Elder, J.H., Prince, S.J.D., Hou, T., Sizintsev, M., Olevskiy, E.: Pre-attentive and Attentive Detection of Humans in Wide-field Scenes. International Journal of Computer Vision 72, 47–66 (2007)
6. Alberto, D.B., Federico, P.: Towards On-line Saccade Planning for High-resolution Image Sensing. Pattern Recognition Letters 27, 1826–1834 (2006)
7. Turk, M., Pentland, A.: Eigenface for Recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)
8. Parker, J.A., Kenyon, R.V., Troxel, D.E.: Comparison of Interpolating Methods for Image Resampling. IEEE Transactions on Medical Imaging 2, 31–39 (1983)
9. Kim, D.H., Lee, J.Y., Yoon, H.S., Cha, E.Y.: A Non-Cooperative User Authentication System in Robot Environments. IEEE Transactions on Consumer Electronics 53, 804–810 (2007)
10. Claude, E.D.: Lanczos Filtering in One and Two Dimensions. Journal of Applied Meteorology 18, 1016–1022 (1979)

Shifting Primes on OpenRISC Processors with Hardware Multiplier

Leandro Marin[1], Antonio J. Jara[2], and Antonio Skarmeta[2]

[1] Department of Applied Mathematics
[2] Research Institute for Oriented ICT (INTICO)
Computer Sciences Faculty, University of Murcia
Reg. Campus of Int. Excellence "Campus Mare Nostrum"
Murcia (Spain)
{leandro,jara,skarmeta}@um.es

Abstract. Shifting primes have proved its efficiency in CPUs without hardware multiplier such as the located at the MSP430 from Texas Instruments. This work analyzes and presents the advantages of the shifting primes for CPUs with hardware multiplier such as the JN5139 from NXP/Jennic based on an OpenRISC architecture. This analysis is motivated because Internet of Things is presenting several solutions and use cases where the integrated sensors and actuators are sometimes enabled with higher capabilities. This work has concluded that shifting primes are offering advantages with respect to other kind of primes for both with and without hardware multiplier. Thereby, offering a suitable cryptography primitives based on Elliptic Curve Cryptography (ECC) for the different families of chips used in the Internet of Things solutions. Specifically, this presents the guidelines to optimize the implementation of ECC when it is presented a limited number of registers.

1 Introduction

Internet of Things proposes an ecosystem where all the embedded systems and consumer devices are powered with Internet connectivity, distributed intelligence, higher lifetime and higher autonomy. This evolution of the consumer devices to more connected and intelligent devices is defining the new generation of devices commonly called "smart objects".

Smart objects are enabled with the existing transceivers and CPUs from the Wireless Sensor Networks (WSNs), i.e. CPUs highly constrained of 8 and 16 bits such as ATMega 128, Intel 8051, and MSP430 [7]. But, since the level of intelligence and required functionality is being increased, some vendors are powering the consumer devices with CPUs not so constrained such as ARM 5 used in the SunSpot nodes [8] from Oracle Lab or the NXP/Jennic JN5139 used in the imote and recently in the first smart light from the market based on 6LoWPAN presented by GreenWave [9].

These smart objects require a suitable security primitives to make feasible the usage of scalable security protocols for the application layer such as DTLS, which

K. Mustofa et al. (Eds.): ICT-EurAsia 2013, LNCS 7804, pp. 540–549, 2013.

has been considered the security to be applied over the Constrained Application Protocol (CoAP) [10] over IPv6 network layer [11].

Specifically, CoAP and the Smart Energy profile for ZigBee alliance (SE 2.0) are considering DTLS 1.2 described in the RFC6347 [12]. This extends the ciphersuites to include the supported by hardware in the majority of the Wireless Sensor Networks transceivers, i.e. AES-128 in CCM mode. In addition, this includes Elliptic Curve Cryptography (ECC) for establishing the session.

Therefore, the challenge is in offering a suitable ECC implementation for the authentication and establishment of the sessions through algorithms such as DTLS.

ECC implementations have been optimized in several works of the state of the art. For example, it has been optimized for constrained devices based on MSP430 in our previous works. But, such as described, the market is not limited to these highly constrained devices therefore it needs to be evaluated how the special primes considered for very specific CPU architectures and conditions are performing for other CPUs.

This work presents the shifting primes and describes how they were used for the MSP430, then it is described the new architecture from the JN5139 CPU, and finally how the shifting primes continue being interesting for the implementation over this higher capabilities, in particular shifting primes offer a feature to carry out the reduction modulo p at the same time that it is carried out the partial multiplication in order to optimize the usage of the registers and consequently reach an implementation which is presenting the best performance from the state of the art for the JN5139 CPU.

2 Shifting Primes

Shifting primes are a family of pseudo-mersenne primes that were designed, in [4], to optimize the ECC cryptography primitives when the CPU is not supporting a hardware multiplication. This type of constrained CPUs is the commonly used in sensors and actuators for home automation and mobile health products. For example, the low category of the MSP430 family from Texas Instrument [7].

Similar primes to the shifting primes have been previously mentioned in [14], but they did not exploited its properties and applications. These new properties which are the used to optimize the implementation for constrained devices without hardware multiplier were described in [4]. In addition, this work presents new features for its optimization in CPUs with hardware multiplication support.

A shifting prime p is a prime number that can be written as follows: $p = u \cdot 2^{\alpha} - 1$, for a small u. In particular we are using for the implementations $p = 200 \cdot 2^{8 \cdot 19} - 1$. There are more than 200 shifting primes that are 160-bit long. The details about this definition can be seen in [4] and [5].

For the implementation of the ECC primitives is used the Montgomery representation for modular numbers. Thereby, computing $x \mapsto x/2(p)$ is very fast even without a hardware multiplier when the shifting primes are used.

Operations using shifting primes can be optimized computing $x \mapsto x/2^{16}(p)$ instead of shifting one by one each step during the multiplication. By using this technique, MSP430 can make a single scalar multiplication within 5.4 million clock cycles in [3].

But, the situation is rather different when the CPU supports hardware multiplication. For this situations, the use of the hardware multiplier through the offered instructions set performs better, since blocks of several bits can be multiplied within a few cycles, for example blocks of 16 bits for a CPU of 32 bits with a 16 bits multiplier such as the located at the JN5139 CPU from Jennic/NXP.

The following sections present as the implementation of the ECC primitives can be optimized for CPUs with hardware multiplier and the advantages that the shifting primes are offering for these high capability CPUs yet.

3 C and Assembler in JN5139

The ECC primitives implementation has been optimized for Jennic/NXP JN5139 microcontroller. The implementation is mainly developed in C, but there are critical parts of the code that require a more precise and low level control, and they have required the use of assembler. In particular, assembler has been used for the basic arithmetic (additions, subtractions and multiplications modulo p).

The target architecture of this chip is based on the OpenRISC 1200 instruction set, and it has been named "Beyond Architecture" or "ba". In particular, the basic instruction set for JN5139 is called "ba1" and the one for JN5148 is called "ba2".

Some of the characteristics that are important in our implementation are:

1. 32 general purpose registers (GPRs) labeled r0-r31. They are 32 bits wide. Some of them are used for specific functions (r0 is constantly 0, r1 is the stack pointer, r3-r8 are used for function parameters and r9 is the link register). See [6, Section 4.4] and [6, Subsection 16.2.1].
2. All arithmetic and logic instructions access only registers. Therefore, they require the use of load and store instructions to access memory. For this purpose, OpenRISC offers the required instructions to load and store in a very flexible way, i.e. this offers operations for bytes, half words and words between registers and memory. The memory operations consume a low number of cycles. Load operations require two clock cycles and store operations require one clock cycle, when there is not cache line miss or DTLB miss. See [1, Page 15].
3. Addition operation offers different instructions; addition with carry and addition with immediate value (`l.add`, `l.addc`, `l.addi`), which consume one clock cycle.
4. Multiplication instruction `l.mul` requires tree clock cycles. See [1, Table 3.2]. This multiplies two registers of 32 bits, but this only stores the result in a single register of 32 bits. Therefore, in order to now loss information, it is only used the least significant part of the registers, i.e. 16 effective bits from the 32 bits, to make sure that the result fit into a single register of 32 bits.
5. The clock frequency of the JN5139 CPU is 16MHz.

Since the RISC principles from the CPU used, the instruction set is reduced, but this offers a very fast multiplication (within 3 clock cycles) with respect to the 16 bits emulated multiplication available in the MSP430 CPUs, which requires more than 150 cycles. Therefore, the support for the multiplication is highly worth for the modular multiplication performance. Such as mentioned, the multiplication offered by the JN5139 CPU is limited to the least significant 32 bits of the result, therefore this requires to limit the multiplication for its usage in the modular multiplications. This is important because the implementation carries out 16x16 multiplications to avoid information loss.

Another important characteristic is that there are available a high number of registers. Therefore, this allows to keep all the information in registers during the multiplication process, and consequently reduce the number of memory operations.

The following sections presents how the multiplication modular is implemented over the JN5139 CPU and how the shifting primes are optimized this implementation thanks to its suitability for the reduction modulo p, in order to reduce the total number of required registers.

4 Multiplication Algorithm

There are different options to compute the product $a \cdot b$ modulo p. The choice of one of them depends on the instruction set and the number of registers available. There are a lot of C implementations that could be optimal for some architectures, but they are rather inconvenient for other architectures.

The decisions considered for this implementation are dependent on this particular architecture, in terms such as the mentioned issues with the multiplication instruction, `l.mul`, which offers a multiplication of two 32-bit numbers, but only offers the least significant 32 bits as a result.

Let x and y, the two operands for the multiplication stored in 16 bits blocks with big endian memory. Then, $x = \sum_{i=0}^{10} x_i 2^{16(10-i)}$ and $y = \sum_{i=0}^{10} y_i 2^{16(10-i)}$. The basic multiplication algorithm requires to multiplicate each $x_i y_j$ and add the partial result from each partial multiplication to the accumulator.

The result from the multiplication of the $x_i y_j$ blocks is a number of 32 bits, which needs to me added to the accumulator. This addition to the accumulator requires previously the shifting of the result to the proper position regarding the index from the $x_i y_j$ blocks, i.e., $m_{ij} = x_i y_j 2^{16(10-i)} 2^{16(10-j)} = x_i y_j 2^{16(10-i-j)}$ requires a shifting of $2^{16(10-i-j)}$ to be added to the accumulator.

Since the result from the multiplication of the $x_i y_j$ blocks is a number of 32 bits, the result can be directly added only when i and j presents the same parity. This is mainly caused because m_{ij} will be divisible by 2^{32}, and consequently it will be aligned to a word (32 bits). Otherwise, when i and j are not presenting the same parity, m_{ij} will be divisible by 2^{16} and not by 2^{32}, consequently the memory is not aligned and it cannot be operated.

A solution for the presented problem with the memory is realign the results m_{ij} when $i + j$ is odd, but this requires to shift operations, and the addition to the accumulator of both results. This means 4 instructions instead of the 1 instruction when the result is . Note, that the 50% of the multiplications will not be aligned.

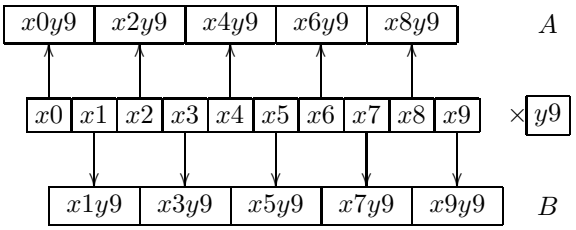
The proposed solution in this work in order to avoid the mentioned extra costs and impact in the performance is to define a second accumulator. Therefore, an accumulator is used for the aligned additions, i.e. $k{\cdot}2^{32}$, and another accumulator is used for the not aligned additions, i.e. $k \cdot 2^{32} + 2^{16}$. Thereby, the realign is only required at the end, when both accumulators are combined.

The proposed solution presents the inconvenient of requiring a high number of registers to store the second accumulator. In the particular case, when the numbers have a length equal to 160 bits, the accumulator needs to be equal to 320 bits, in order to store the result from 160x160. Therefore, it is required 10 registers for a single accumulator, and consequently 20 registers for the two required accumulators. In addition, one of the operands needs to be also stored into the registers, i.e. 160 bits stored in 10 registers, 16 bits per register, note that it is stored only 16 bits per register even when it is feasible to store 32 bits, since the previously described limitations by the hardware multiplication. In summary, it is required 30 registers to keep into the registers the accumulators and one of the operands, but it is not feasible because additional registers for the temporal are required.

But, a solution is feasible to maintain both accumulators into the registers and at the same time the additionally required registers thanks to the features from the shifting primes.

Shifting primes allows to carry out the reduction modulo p, while it is added the partial result to the accumulator. Thereby, the result is 160 bits (when p is a 160 bits number). This allows to keep the two accumulators of 160 bits, instead of the previously mentioned 2 accumulators of 320 bits. Therefore, only 5 registers are required for each accumulator, i.e., 10 registers for both accumulators, what is feasible.

Let two accumulators A and B, $A = \sum_{i=0}^{5} A_i 2^{32 \cdot i}$, $B = \sum_{i=0}^{5} B_i 2^{32 \cdot i}$. A is used for the aligned operations and B for the not aligned operations. Let the previously mentioned operands x and y. The first multiplication iteration is to multiplicate the first operand by $x9$, and after add it to the adequate accumulator such as follows:



Then, it is established the operand x into the registers r13 ,..., r20, the accumulator A in r22 ,..., r26 and the accumulator B in r27 ,..., r31. The assembler

code required to carry out this partial multiplication is composed by 10 l.mul, where each one requires 3 cycles.

```
l.mul  r31,r20,r3        l.mul  r24,r17,r3
l.mul  r26,r21,r3        l.mul  r28,r14,r3
l.mul  r30,r18,r3        l.mul  r23,r15,r3
l.mul  r25,r19,r3        l.mul  r27,r12,r3
l.mul  r29,r16,r3        l.mul  r22,r13,r3
```

The next step is to multiplicate by $x8$, and shift 16 bits the result. Note, that it should be required to shift 16 bits the accumulator, but instead of that operations, the proposed solution stores this partial result in the accumulator B. The global effect is a 16 bits in the accumulator at the end. For the next multiplication, it is required to shift the accumulator A by a block, i.e. 32 bits. For this operations, it is taking into account that $p = \texttt{0xc800} \cdot \texttt{0x10000}^9 - 1$ and consequently $1 \equiv \texttt{0xc800} \cdot \texttt{0x10000}^9$ modulo p, therefore it is equal to:

$$A = A_0 \cdot \texttt{0x10000}^8 + A_1 \cdot \texttt{0x10000}^6 + A_2 \cdot \texttt{0x10000}^4 + A_3 \cdot \texttt{0x10000}^2 + A_4 =$$

$$A_0 \cdot \texttt{0x10000}^8 + A_1 \cdot \texttt{0x10000}^6 + A_2 \cdot \texttt{0x10000}^4 + A_3 \cdot \texttt{0x10000}^2 + A_4 \texttt{0xc800} \cdot \texttt{0x10000}^9$$

It can be splitted into two block of 16 bits, then $A_4 = A_4^H \cdot \texttt{0x10000} + A_4^L$, and it can be considered:

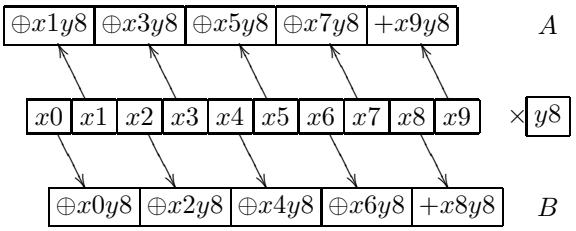
$$A = (\texttt{0xc800} \cdot A_4^H \cdot \texttt{0x10000}^8 + A_0 \cdot \texttt{0x10000}^6 + A_1 \cdot \texttt{0x10000}^4$$
$$+ A_2 \cdot \texttt{0x10000}^2 + A_3) \cdot \texttt{0x10000}^2 + \texttt{0xc800} \cdot A_4^L \cdot \texttt{0x10000}^9$$

The value $\texttt{0xc800} \cdot A_4^L \cdot \texttt{0x10000}^9$ is moved to B, since now it has the adequate alignment to be combined with the other accumulator. Then, the accumulator A can be shifted 32 bits with only change the role of the register. The change of the role for the registers does not require any explicit instruction, it is only a programming issue, which can be adapted when the loop from the multiplication is unrolled. Therefore, it is unrolled the 10 iterations of the loop (one iteration for each 16 bits from the 160 bits of the operand).

Therefore, the registers rotation is directly programmed in the code. Note that the register r6 has the value 0xc800 during all the operation. The code is as follows:

```
l.andi  r8,r31,0xffff        l.add   r26,r26,r8
l.mul   r8,r8,r6             l.addc  r8,r7,r0
l.srli  r31,r31,16           l.slli  r8,r8,16
l.mul   r31,r31,r6           l.add   r31,r31,r8
```

After the change is carried out in the accumulator, then it is carried out the multiplication with the block x_8, and it is added to the appropriated accumulator. The scheme is as follows (where + represent the addition operation, and $\oplus$ the addition with carry):



Finally, in terms of assembly code is:

```
l.mul  r7,r20,r3          l.mul  r8,r21,r3
l.add  r25,r25,r7         l.add  r31,r31,r8
l.mul  r7,r18,r3          l.mul  r8,r19,r3
l.addc r24,r24,r7         l.addc r30,r30,r8
l.mul  r7,r16,r3          l.mul  r8,r17,r3
l.addc r23,r23,r7         l.addc r29,r29,r8
l.mul  r7,r14,r3          l.mul  r8,r15,r3
l.addc r22,r22,r7         l.addc r28,r28,r8
l.mul  r7,r12,r3          l.mul  r8,r13,r3
l.addc r26,r26,r7         l.addc r27,r27,r8
l.addc r7,r0,r0
```

The last carry values from both accumulators needs to be added. The carry of A is shifted 16 bits and added to the most significant part of B. The final carry from this operation is added to the carry of B.

```
l.addc r8,r0,r0           l.add  r26,r26,r8
l.slli r8,r8,16           l.addc r7,r7,r0
```

Now, the role of the accumulators A and B are interchanged, with a shifting of 32 bits in A. The pending carry is added to the most significant part of A.

The presented process needs to be repeated 9 times until that it is completed all the blocks of the operand, i.e. y_i. At the end, all the accumulators are added in order to get the final result.

5 Results and Evaluation

The evaluation is focused on the multiplication modulo p, since it is the most critical part in the ECC algorithms. There is a very rich literature about how to implement ECC primitives by using the basic modular arithmetic. For this purpose, it needs to be fixed a curve and implement the point arithmetic. A wide variety of curves, point representations and formulas for point addition and point doubling can be found in [2].

For the prime $p = 200 \cdot 256^{19} - 1$ we have chosen the Weierstrass curve $y^2 = x^3 - 3x - 251$. The number of points of this curve is

$p + 1257662074940094999762760$, that is a prime number. We have chosen the parameter -3 to use the formulas for point addition and point doubling in Jacobian coordinates given in [2].

The time that we have considered as a reference is the time required for a key generation. This requires the selection of a random number s_K (the private key) and the computation of the scalar multiplication $[s_k]G$ where G is a generator of the group of points in the curve. The generator has been set up with the following coordinates: $x_G =$ `0x9866708fe3845ce1d4c1c78e765c4b3ea99538ee` and $y_G =$ `0x58f3926e015460e5c7353e56b03dd17968bfa328`

The time required for the scalar multiplication is usually computed in terms of the time **M** required for a single modular multiplications. A standard reference is [15] that gives 1610**M** for 160-bit scalar multiplication. This result requires some precomputations and considers that computing a square is a bit faster than a standard multiplication, 0.8**M**. We have made a rather optimized multiplication with a code that requires around 2Kb. To have another function for squaring would require more or less the same and all the other precomputations could increase the size of the program too much. We have used an implementation that requires 2100**M** (1245 multiplications and 855 squares).

Standard literature ignores the time required for other operations different from modular multiplication. We have computed in our case that the key generation spends 83.13% of the time doing modular multiplications, this is a big percentage, but not all the time. Of course, our biggest efforts have been done in the optimization of this operation.

The following table presents the real time required for the key generation, one single modular multiplication, and finally for 2100 of them.

Key Generation	140.37 ms
Single multiplication	54.9 μs
2100 multiplications	115.29 ms
rest	25.08 ms

Since the CPU clock from the JN5139 is equal to 16MHz, 54.9 μs are 878 clock cycles in real time. The number of cycles for reading the code is around 750 cycles, this is the theoretical number of cycles. But, the real time is a little bit higher than the theoretical time because the external interruptions, cache failures, and the pipeline could require some extra cycles to execute the code.

6 Conclusions and Future Work

The first conclusion is that the multiplication algorithm presents a high dependence to the CPU architecture where it will be evaluated. From our experience, we have experimented with the MSP430 architecture, which is based on a 16 bits microcontroller, without support for hardware multiplication, 16 registers and a limited set of instructions. For this work, it has been evaluated an architecture

with higher capabilities, specifically, an architecture based on OpenRISC with 32 bits operations, support for hardware multiplication, 32 registers with a very low cost per instruction, and an extended set of instructions. These huge differences make very difficult to compare among different implementations, at least that they have tested over the same architecture and exploiting the same features. For example, it can be found an implementation of ECC for MSP430 and JN5139 in [13] which presents a very low performance, since it is implemented over the WiseBed Operating System. Therefore, even when they are using the same architecture the results are highly different because they are not being able to exploit the main benefits from the architecture.

The second conclusion is that shifting primes have demonstrated be a very useful primes, which are offering a very interesting set of properties in order to optimize the implementations for constrained devices. First, it was optimized for the MSP430 CPU thanks to its low quantity of bits set to 1, which simplifies some iterations from the multiplication when it is implemented through additions and shifts as a consequence that this was not supported supported by the hardware. Second, it has been also presented how to exploit the shifting primes for architectures, such as the JN5139 based on OpenRISC, which are supporting hardware multiplication. The optimization for the JN5139 has been focused on its suitability for the reduction modulo p, while it is added the partial result to the accumulator in order to make feasible the exploitation of the hardware multiplication over the available registers.

Finally, it needs to be considered hybrid scenarios, since in the different use cases from the Internet of Things, it will be common to find a same solution with multiple CPUs in the sensors, actuators and controllers. For example, it could be found a JN5139 module in the controller, since this has a higher memory and processing capabilities to manage multiple nodes requests and maintenance. However, the most common CPU for the sensors and actuators will be the MSP430, since this presents a lower cost but yet enough capabilities for their required functionality. Therefore, it is very relevant to have this kind of primes and implementations such as the presented in this work, which is feasible for devices with different capabilities. For that reason, our future work will be focused on demonstrate a scenario where MSP430 and JN5139 are integrated into the same solution, and both implement high level security algorithms such as DTLS for CoAP, using both of them certificates built with shifting primes-based keys. making thereby them feasible to interoperate and exploit the described optimization.

Acknowledgment. This work has been carried out by the excellence research group "Intelligent Systems and Telematics" granted from the Foundation Seneca (04552/GERM/06). The authors would like also thanks to the Spanish Ministry of Science and Education with the FPU program grant (AP2009-3981), the Ministry of Science and Innovation, through the Walkie-Talkie project (TIN2011-27543-C03-02), the STREP European Projects "Universal Integration of the Internet of Things through an IPv6-based Service Oriented Architecture enabling

heterogeneous components interoperability (IoT6)" from the FP7 with the grant agreement no: 288445, "IPv6 ITS Station Stack (ITSSv6)" from the FP7 with the grant agreement no: 210519, and the GEN6 EU Project.

References

1. OpenRISC 1200 IP Core Specification (Preliminary Draft), v0.13 (2012)
2. Bernstein, D.J., Lange, T.: Explicit-formulas database, http://hyperelliptic.org/EFD
3. Marin, L., Jara, A.J., Skarmeta, A.F.G.: Shifting primes: Optimizing elliptic curve cryptography for 16-bit devices without hardware multiplier (preprint)
4. Marin, L., Jara, A.J., Skarmeta, A.F.G.: Shifting Primes: Extension of Pseudo-Mersenne Primes to Optimize ECC for MSP430-Based Future Internet of Things Devices. In: Tjoa, A M., Quirchmayr, G., You, I., Xu, L. (eds.) ARES 2011. LNCS, vol. 6908, pp. 205–219. Springer, Heidelberg (2011)
5. Marin, L., Jara, A.J., Skarmeta, A.F.G.: Shifting primes: Optimizing elliptic curve cryptography for smart things. In: You, I., Barolli, L., Gentile, A., Jeong, H.-D.J., Ogiela, M.R., Xhafa, F. (eds.) IMIS, pp. 793–798. IEEE (2012)
6. opencores.org. OpenRISC 1000, Architecture Manual, Rev. 2011-draft4 (2011)
7. Davies, J.H.: MSP430 Microcontroller Basics, 9780080558554. Elsevier (2008)
8. Castro, M., Jara, A.J., Skarmeta, A.: Architecture for Improving Terrestrial Logistics Based on the Web of Things. Sensors 12, 6538–6575 (2012)
9. Hoffman, L.: GreenWave Reality Announces Partnership with NXP, GreenWave Reality (2011), http://www.greenwavereality.com/greenwave-reality-announces-partnership-with-nxp-semiconductors/
10. Shelby, Z., Hartk, K., Borman, C., Fran, B.: Constrained Application Protocol (CoAP), Internet-Draft. Internet Engineering Task Force (IETF) (2012)
11. Jara, A.J., Zamora, M.A., Skarmeta, A.F.: GLoWBAL IPv6: An adaptive and transparent IPv6 integration in the Internet of Things. Mobile Information Systems 8(3), 177–197 (2012)
12. Rescorla, E., Modadugu, N.: RFC6347 - Datagram Transport Layer Security Version 1.2, Internet Engineering Task Force (IETF) (2012) ISSN: 2070-1721
13. Chatzigiannakis, I., Pyrgelis, A., Spirakis, P.G., Stamatiou, Y.C.: Elliptic Curve Based Zero Knowledge Proofs and their Applicability on Resource Constrained Devices. In: 2011 IEEE 8th International Conference on Mobile Adhoc and Sensor Systems (MASS), pp. 715–720 (2011), doi:10.1109/MASS.2011.77
14. Hasegawa, T., Nakajima, J., Matsui, M.: A Practical Implementation of Elliptic Curve Cryptosystems over GF(p) on a 16-bit Microcomputer. In: Imai, H., Zheng, Y. (eds.) PKC 1998. LNCS, vol. 1431, pp. 182–194. Springer, Heidelberg (1998), doi:10.1007/BFb0054024
15. Cohen, H., Miyaji, A., Ono, T.: Efficient Elliptic Curve Exponentiation Using Mixed Coordinates. In: Ohta, K., Pei, D. (eds.) ASIACRYPT 1998. LNCS, vol. 1514, pp. 51–65. Springer, Heidelberg (1998)

Author Index

Abbasi, Sedigheh 447
Abu, Nur Azman 519
Alam Aldeen, Nisreen 441
Alhasa, Kemal Maulana 214
Armando, Alessandro 469
Asfand-e-yar, Muhammad 233
Aziz, Noraniah Abdul 162

Baba, Kensuke 528
Bagchi, Susmit 81
Biuk-Aghai, Robert P. 31
Bkakria, Anis 426
Breier, Jakub 285
Bui, Thach V. 324

Castiglione, Aniello 469
Cattaneo, Giuseppe 507
Chandra, Dissa R. 41
Chen, Hua 318
Chen, Xiaofeng 243, 373
Choi, Dongjin 253
Cimato, Stelvio 447
Costa, Gabriele 469
Cuppens, Frédéric 426
Cuppens-Boulahia, Nora 426

Damiani, Ernesto 447
Dang, Tran Khanh 11, 101, 111, 121
Dang, Tran Tri 121
Dang, Van H. 324
De Maio, Giancarlo 507
Dhar, Aritra 141
Duong, Anh-Duc 409
Dutta, Ratna 496

Egawa, Serina 528
Ekaputra, Fajar Juang 71
Ernawan, Ferda 519

Faruolo, Pompeo 507
Feng, Yansheng 243
Fernandez, José M. 426
Ferraro Petrillo, Umberto 507
Fiore, Ugo 469
Fudholi, Dhomas Hatta 223

Gao, Jing 173
Georgescu, Adela 353
Govindaraju, Rajesri 41

Hadining, Aulia F. 41
Hagihara, Shigeki 60
Hasibuan, Zainal Arifin 192
He, Mingxing 513
Hendrik 223
Herawan, Tutut 162
Hou, Shuhui 312
Huang, Minhuan 306, 318
Huang, Yi-Li 392
Hudec, Ladislav 285

Jara, Antonio J. 540
Jensen, Jostein 343
Jia, Wentao 337

Kaneko, Tomoko 331
Kang, Ho-Seok 490
Kim, Pankoo 253
Kim, Sung-Ryul 490
Ksiezopolski, Bogdan 261
Kuang, Xiaohui 479

Le, Thu Thi Bao 111
Leu, Fang-Yie 392
Li, Hui 300, 383, 459
Li, Jin 373
Li, Rui 337
Li, Xiang 479
Li, Xiao 513
Li, Xiaoqing 383
Liu, Beishui 459
Liu, Bo 453
Liu, Jung-Chun 392
Liu, Lingxia 306
Liu, Tingting 459
Lu, Huabiao 271

Ma, Hua 243, 373
Marin, Leandro 540
Merlo, Alessio 469
Mitra, Sarbari 496
Moon, Hae-Min 534

Mukhopadhyay, Sourav 496
Muljo, Hery Harjono 202
Mustofa, Khabib 21

Nam, Nguyen Hoai 183
Ng, Liang Shen 162
Ngo, Chan Nam 101
Nguyen, Huynh Tuong 183
Nguyen, Nhung T.H. 324
Nguyen, Oanh K. 324
Nguyen, Thi Ai Thao 11
Nguyen, Thuc D. 324
Nguyen, Quang-Hung 183
Nien, Pham Dac 183
Nishide, Takashi 363

Olimid, Ruxandra F. 399

Pan, Sung Bum 534
Pardamean, Bens 202
Pardede, Eric 223
Park, Youngho 416

Quirchmayr, Gerald 441

Rahayu, Wenny 223
Ramdhani, Neila 152
Ren, Yizhi 295
Rhee, Kyung-Hyune 416, 502
Rusinek, Damian 261

Sahib, Shahrin 519
Sakurai, Kouichi 363
Sarkar, Pinaki 141
Sasaki, Ryoichi 312
Selviandro, Nungki 192
She, Yuchao 300
Shimakawa, Masaya 60
Sir, Yosua Albert 21
Skarmeta, Antonio 540
Su, Jinshu 271
Sulistyo, Selo 50
Sun, Xiaoxia 318
Sunindyo, Wikan Danar 71
Suparta, Wayan 214
Suprapto 91
Sur, Chul 416
Suryana, Nanna 519
Syafar, Faisal 173
Syalim, Amril 363

Takaya, Mayumi 131
Tanaka, Hidehiko 331
Thoai, Nam 183
Tjoa, A Min 233
Tran, Minh-Triet 409
Truong, Quynh Chi 121
Truong, Toan-Thinh 409
Tsuruta, Yusuke 131

Uehara, Tetsutaro 312

Venkatesan, Hari 31
Verderame, Luca 469

Wahid, Fathul 1
Wang, Dongxia 306
Wang, Huan 513
Wang, Xiaofeng 271
Wardoyo, Retantyo 91
Wei, Guanghui 459
Wen, Huaqing 502
Wen, Yan 485
Wierzbicki, Adam 261
Wiradhany, Wisnu 152
Wu, Beibei 295
Wu, Chunqing 453

Xie, Xingxing 373
Xu, Jian 295
Xu, Ming 295

Yamamoto, Shuichiro 331
Yamamura, Akihiro 131
Yang, Jing-Hao 392
Yiu, Siuming 312
Yonezaki, Naoki 60
You, Ilsun 253, 453, 469
Yu, Wanrong 453

Zhang, Chunyan 337
Zhang, Haiping 295
Zhang, Yinghui 383
Zhao, Baokang 271, 453
Zhao, Jinjing 306, 318, 479, 485
Zhao, Lu 490
Zheng, Ning 295
Zhu, Hui 243, 300, 383, 459
Zou, Dingjie 453